Understanding What affects Career Progression Using LinkedIn and Twitter Data

Yiming Pan, Xuefeng Peng, Tianran Hu, and Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, USA
{*ypan17, xpeng4, thu, jluo*}*@cs.rochester.edu*

***Abstract*—Nowadays competitions in workplaces become increasingly intense. People are looking for promotion strategies that could help them move up faster in their career paths. Our main objective is to determine how factors such as personality, industry and education background impact one's career path, and the highest career stage one could reach. In this study, we bring a novel methodology to determine a career stage based on the job title and company information, so that a career path that consists of several stages could represent the occupational growth. We associate individuals' career paths with their education backgrounds, unique thinking styles, interests, and personalities by analyzing extensive users from Social Media. Our study shows that those able to move up faster and higher share particularly similar traits, characteristics and tendencies. Finally, we employ machine learning techniques to predict career progression with a promising accuracy.**

***Keywords*-Career Prediction; Promotion; Personality; Education; Big Five Trait; Social Media;**

I. Introduction

Many studies on sociology and psychology reveal that many aspects of people's lives are deeply connected with their jobs [1], [2]. While most of the previous work focus on job types [3], the employment status and its impact [4], [5], [6]. Few study focused on career progression, and the factors that impact people's career paths. In this paper, we associate career paths with educational backgrounds and personal traits which include personality and lifestyle. A career path is described as eight career stages and each stage contains a set of equivalent job titles, so that it represents people's occupational growth in their career. We study people's distinct characters, emotions and interests along with their unique career paths.

With the accelerating development of social media, individuals would like to share thoughts and personal opinions by sending posts on social networks like Twitter[1]. Social medial posts have been proven to be an effective information source in gaining knowledge of people [7], [8]. Therefore, we collect Twitter posts to analyze users' personality, emotions and interests. We apply Receptiviti API[2] to measure these traits. It relies on the Linguistic Inquiry and Word Count (LIWC 2015)[3] and it gives us Thinking Style Insights, Big Five Traits, Social Style Insights, Emotional Style Insights, Working Style Insights, Interests and Orientations. These insights and traits help describe personality [9] and are clearly interpreted in the Personality Measurement section later.

As LinkedIn[4] is a business oriented social network and LinkedIn users would like to post their career and education experience, we gather job and education information with timestamps from LinkedIn. Although we could directly utilize job titles from their profiles, those job titles in different companies don't necessarily represent the same career stage. For example, getting promoted to a CEO in a large enterprise like Google is far more difficult than in a small company. To come up with more accurate career stages, we evaluate a job title based on the size of its company, which could be referred from the number of employees. Thus, we collect company sizes from LinkedIn homepages of these companies and divide them into four scales: startup, small business, medium business, enterprise. Furthermore, since the job titles differ in different industries, especially the job positions in management level. For example, 'chancellor' is one of the highest position in school management, but it only comes up in an academic or education career. Hence, we also take the industry of an individual's career into consideration. Finally, we match job titles to career stages separately for each company scale and each industry, which is interpreted in detail in the Career Stage Partition section.

Interesting and significant correlations between personal traits and career progression are uncovered. For example, we observe that people who move up fast in their careers talk a lot about their bodies and food, which suggests that they pay more attention to their daily lives and it might help their careers. According to our results, for each industry and each career stage, there are distinct traits helping people move up faster. For example, in a large enterprise, openness helps people get to upper management level faster, but has little impact on the progression at the entry level. Based on these findings, we employ Support Vector Regression and Ensemble Learning model to predict the time of employee's future promotions. A promising accuracy indicates that it is practical to utilize human characteristics for career progres-

[1]http://www.twitter.com
[2]http://www.receptiviti.ai
[3]http://liwc.wpengine.com
[4]http://www.linkedin.com

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

sion prediction tasks.

We summarize our contributions as follows:

- We propose a novel methodology to evaluate people's career stages from job titles and construct career paths.
- We explore significant factors that impact careers across industries and career stages.
- We develop a promotion predictor with high accuracy and demonstrate the potential of our career progression model.

II. Related Work

It is worth mentioning that several researches have been dedicated to occupation analysis from social networks [3], [10], [11]. Preoctiuc inferred a user based upon user profiles and social contents [11]. However, most of them used the information from one single source, which makes it difficult to comprehensively characterize a user's personality from various aspects. The author of [3] predicted the occupation of a user by analyzing the Big Five Traits from social media posts.

Besides occupation analysis, the relationships between personality and job have been investigated in several aspects. The correlation between Big Five Traits and job criteria is explored [12], [13], which suggests that Conscientiousness and Neuroticism are valid predictor for job performance. But their study did not provide any valid approach to evaluate job performance, by contrast we extract career paths and utilize promotion records to evaluate career performance. The authors of [14] investigate the correlation between job satisfaction and Big Five Traits, and they found that job satisfaction is positively correlated with Extraversion, Agreeableness and Conscientiousness, but negatively correlated with Neuroticism. There is also work focusing on specific job types. Perceived social support, job stress, health, and job satisfaction among nurses are studied in [15]. [16] compares the U.S. hospitality industry employees with other industries on work attributes, demographics, and class perceptions. The results show that the people of this industry tend to be less satisfied with the job, and take their work as unimportant elements in their self-accomplishments. While these works do not focus on promotions or career path prediction, which is one of the most critical aspects that people care about.

There is also work related to studying people's career paths. The authors of [17] presented a multi-source learning framework which integrated social network sources and modeled a career path. However, what they extracted to represent a user are demographic features, LIWC features and user topic features. The authors did not focus on human characteristics like personality, emotions or interests.

Although some work has been done, previous work focus on occupation analysis, work status or job satisfaction. Hence, it is still not clear what kind of people are favored for promotions and what unique traits could help speed up career progression. Moreover, due to the inherent limits of transitional data collection methods, previous work of sociology and psychology usually suffers from the problem of small sample size. In our work, a large dataset is collected from social media platforms, and we compute multiple human characteristics to investigate the consequential traits across career stages and job categories.

III. Data Acquisition

In this section, we mainly interpret our dataset and discuss about personality measurement, as well as the career stage partition. As we would like to collect both twitter posts and LinkedIn profiles from the same users, we utilize a valid dataset provided by the author of [3], which collected tweets and job information of around 150 thousand user profiles. The author collected these data from about.me[5], which is an integration platform allows users to link multiple online identities such as LinkedIn and Twitter. The fact that those users link their identities from different social media platforms help ensure veracity as well as quality of their information. Note that the author randomly sampled users from the most popular 1,000 male and female first names in the U.S[6], which means those users should not have biases towards certain directions. We also remove the users who do not have enough English tweets (less than 300 words), to guarantee the significance of our results. Furthermore, [3] applied Twokenizer to clean tweets and categorized jobs by clustering skills that endorsed by others. Instead of directly use the industry tags on their LinkedIn pages, we utilize those clustered job categories as the industries of those users to ensure the precision. Figure 1 represents the distribution of people's industries. In order to analyze the correlation between career progression and personality across different industries, we select top four industries (Information Technology, Marketing, Media and Public Relation) to explore, which will be demonstrated in the Correlation Analysis section. Beyond the LinkedIn profiles, we would like to know how big are the companies that each user works for, so that we could determine the career stages from their job titles. Thus, for each user we go through his or her experience, and for each experience we fetch the company size from the LinkedIn homepage of that company. In this process, we remove users that only have one experience, since we are not able to extract their career paths. Although there exist users who do not diligently report each job transition, the start dates and end dates of each experience should still be reliable. As long as we take all the experiences that they've already updated, it will not influence the data veracity. Finally, we eliminate the users whose company can not be found on LinkedIn, and end up with 9246 users in total. Furthermore, for each user in our dataset, we fetch the education section from his

[5]https://about.me

[6]http://www.namelist.com/

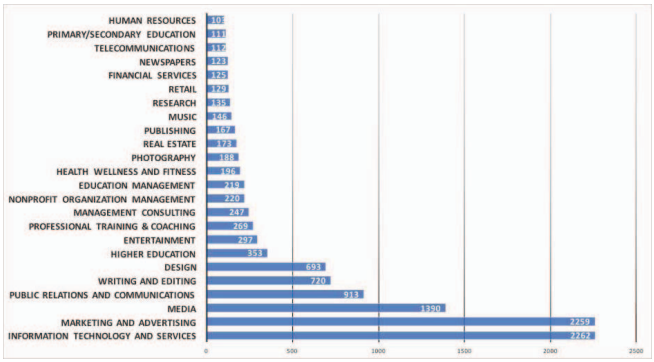

Figure 1: Distribution of people's industries.

or her LinkedIn page. After that, we take the most recent school graduated and look for its rank on U.S. News and World Report. Since there are lots of users don't have the education section and some of the schools are not included in the ranking system, we end up with 3560 records with users' school ranks.

A. Personal Traits Measurement

The way people use words in their daily lives are providing rich information about their beliefs, fears, thinking patterns, social relationships, and personalities. Previous works showed that people's personalities could be computed from their tweets [18]. We find that the Linguistic Inquiry and Word Count tool (LIWC) is a widely tested, validated, and applied system for performing psycholinguistic text analysis[19], [20]. In 2015, a new version of LIWC was released along with Receptiviti, a commercial brand of the tool. In addition to the LIWC categories, Receptiviti includes insights about personality, thinking style, authenticity, and relationships. This service relies on LIWC 2015 with its dictionary composed of almost 6400 words, words stems and select emoticons. [21] validated this tool on Twitter posts, overall the Mean Absolute Error rates (MAE) of Recpetiviti predictions ranges from 15% to 30%. Although the error rate is a bit higher than other personality prediction algorithm, it maintains relative scores between groups of subjects and provides us psychological insights other than the traditional Big Five Traits. We eliminate all the users whose tweets all together have less than 300 words, so that we could maintain the accuracy of the predictions. Then we feed all the tweets to Receptiviti API and get an personal trait vector for each user as follows:

- Thinking Style Insights: thinking style, persuasive, reward bias
- Big Five Traits: openness, conscientiousness, extraversion, agreeableness, neuroticism
- Social Style Insights: social skills, insecure, cold, family orientation
- Emotional Style Insight: adjustment, happiness, depression
- Working Style Insights: independent, power driven, type-A, workhorse
- Interests and Orientations: friendship focus, body focus, health oriented, sexual focus, food focus, leisure oriented, money oriented, religion oriented, work oriented, netspeak

Table I: IT corporations' five job levels and their job titles.

Entry Level	Software Engineer, Software Developer
Lower Management Level	Manager, Supervisor
Middle Management Level	Director
Upper Management Level	Vice President
Executive Level	CEO, CTO, Owner, Founder

Table II: Four company categories by Size.

Company Category 0	<100 employees
Company Category 1	100-999 employees
Company Category 2	1000-9999 employees
Company Category 3	>10000 employees

B. Career Stage Partition

First, we learn the general hierarchy of job titles and investigate the salary and technical requirement for each title. Then we divide all the job titles into five levels, which are entry level, lower management level, middle management level, upper management level and executive level. Each job level consists a set of job titles that require similar prior experience or skills. We label the job level from 0 to 4, where 0 refers to the lowest level and 4 refers to the highest. It is worth noting that job titles vary according to the company and the industry. For example, in an Information Technology company, job titles like Software Engineer or Software Developer refer to the entry level jobs, while in a consulting firm, entry level job titles include Analyst, Consultant, Associate, etc. Table I demonstrates the job level partition and job title examples for each level in an Information Technology corporation. However, even within the same industry, the job title with the same name does not necessarily have the same work and duty, like senior software engineer in IBM is not the same as the senior software engineer in Google. In this study, we consider that these positions are similarly difficult to reach, so any transition at the same job level will not be identified as a career progression. We define $LEVEL(i)$ as a function that matches keywords in the string of a job title and returns the corresponding numerical job level we labeled. We prefer to match keywords like 'director', 'supervisor' or 'engineer' rather than match the exact titles is because the titles share tons of variation and it is not practical to match all the possibilities. Hence the job level numbers $LEVEL(i)$ return vary from 0 to 4.

However, we cannot directly infer the career stage from the job level, as the same job level in companies with different sizes may not be considered as the same career

stage. For example, the entry-level software engineer in an arbitrary software company is hardly to get a job as the entry-level software engineer in Google. As Table II shows, we further categorize companies into four bins by size, based on [22]'s contribution as well as guidance from U.S. Small Business Administration[7]. More specifically, company category 0 consists of startup companies that have less than 100 employees, Company Category 1 consists of small businesses that have 100 to 999 employees, Company Category 2 consists of medium businesses that have 1000 to 9999 employees, Company Category 3 consists of large enterprises that have more than 10000 employees. We use SIZE(N) to evaluate the company category number from 0 to 3 if $N \in [10, 99999]$:

$$SIZE(N) = \begin{cases} \lfloor \log_{10} N \rfloor - 1, & \text{if } 10 \leq N \leq 99999 \\ 0, & \text{if } N < 10 \\ 3, & \text{if } N > 99999 \end{cases}$$

Thus one's career stage come out with the consideration of both the job level of one's job title and the size of one's company:

$$STAGE(i) = LEVEL(i) + SIZE(N)$$

where i is the string of one's job title, N is the number of employees in that particular company.

Next, as Figure 2 shows, we construct a job level mapping system that maps the job levels in different sized company to career stages in a same scale. As Figure 2 shows, the final career stage ranges from 0 to 7, where the lowest stage 0 represents a software engineer at a IT startup and the highest stage 7 corresponds to a CEO at large enterprise like Google. We also consider that job levels at the same career stage are similarly difficult to reach, thus any transition between job positions in the same career stage is not considered as career progression. For example, if a user was a CEO in a small business and he left the job for a vice president in a medium business, then it will not be counted as a progression.

C. Career Progression Formulation

As we have career stages defined, we evaluate the career progression as the duration an individual X works for getting promoted from stage $N-1$ to stage N:

$$DURATION(N) = T_{min}(N) - T_{min}(N-1)$$

where $T_{min}(N)$ refers to the earliest time stamp that user X works at stage N and $T_{min}(N-1)$ refers to the earliest time stamp that the user works at stage $N-1$. For each user, $T_{min}(N)$ returns the start date of the first job at stage N. And for those people who do not have jobs at stage $N-1$, we search for the job at the next highest stage until we reach stage 0. We calculate the duration of the path in total before reaching stage N:

$$PATH(N) = \sum_{1 \leq i \leq 7} DURATION(i)$$

as a more intuitive evaluation of career progression. The purpose of this study is to predict $PATH(N)$, which is the duration from the start date of the first job at the lowest career stage to the start date of the first job at stage N. Therefore, we go through all the experiences of a user in chronological order, if the experience brings the user to a new higher career stage N, then we calculate $PATH(N)$ and construct a new feature vector that consists of personality measurement, industry, stage and duration. We collect all the feature vectors together to be our training set for future correlation analysis and model training process.

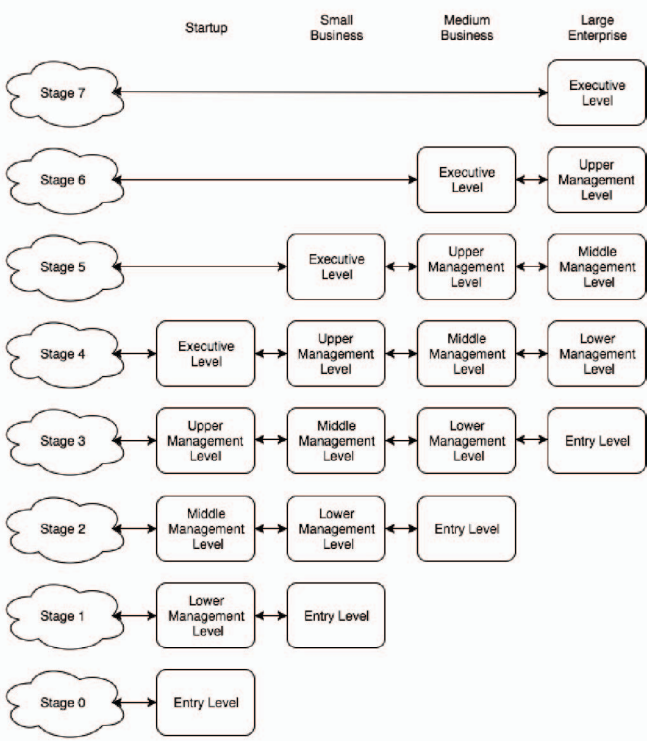

Figure 2: A mapping system between career stages and job levels.

IV. Correlation Analysis

With the career progression well defined, we explore the factors that might impact the career progression, such as industry, education and personality. In different industries, there exist unique traits that help people move up fast. Moreover, the traits that affect peoples' career progression depend on their target stages. For example, openness helps people get to stage 6 faster, but has little impact on the

[7]https://www.sba.gov/

progression towards stage 1. Thus, we analyze the impact of all the traits for each industry and for each career stage.

A. Personal Traits Analysis

First of all, we focus on how personality impacts career progression overall, considering people in all the industries. We put people in different stages all together and we compute the Pearson correlation coefficients for each trait, as shown in the last column of Figure 3. Negative correlation indicates that people who have a relatively faster career progression (shorter duration of staying at the same stage) get higher scores on those traits. The most negatively correlated traits we find are insecure, body focus and food focus. The observation suggests that those people are relatively more cautious and thoughtful when dealing with other people. High score on body focus and food focus reveal that people relatively pay more attention to the quality of their daily lives, including fitness and nutrition which directly related to physical and mental health. While, positive correlations represent that people who have a relatively slower career progression get higher scores on those traits. The most positively correlated features are conscientiousness and adjustment. To our surprise, higher conscientiousness people have slow progression. However, our guess is that although being conservative and disciplined might be favored by superiors, nowadays most companies inspire creativity and encourage innovations. Also, people with fast career progression are not emotionally stable, they are more sensible to the environment. We then analyze the correlations across different stages and different industries.

1) Different Industries: We observe interesting differences across industries, as Figure 3 shows. In each industry exists people that moving up fast with unique traits, which suggest that the factors that impact career progression vary from industry to industry. However, according to the distribution of people's industries shown in Figure 1, we look into top four industries that contain most of the people: Information Technology, Marketing, Media and Public Relation. For the rest industries, they are relatively less popular, thus we decide not to separate them out so as to maintain the accuracy of our findings. Among those most popular industries, we discover interesting facts which are all statistically significant with p-value < 0.05 and summarized as follows:

For IT industry, we have 2008 records describing users that work at either software companies or Internet related companies. The most positively correlated trait is **money oriented** and **conscientiousness**, while the most negatively correlated trait is **insecure**. Those findings indicate that those software engineers who move up fast in the company have a relatively higher score insecure but lower score on conscientiousness and money oriented. The words they use are relatively less conservative but more cautious and introverted, and they talk less about money and finance. Previous work has shown that software engineers are the least agreeable among those four industries and they are a group of people that do not like to compromise [3]. Interestingly, our finding indicates that in IT industry people who have faster career progression are relatively more introverted and self-conscious among all the software engineers.

	IT	MARKET	MEDIA	PR	OVERALL
thinking_style	0.07	0.14	0.13	0.14	0.12
persuasive	0.06	0.15	0.16	0.09	0.13
reward_bias	0.00	0.08	0.01	-0.02	0.02
openness	-0.06	-0.11	0.00	-0.08	-0.07
conscientiousness	0.13	0.26	0.13	0.08	0.16
extraversion	0.04	0.22	0.17	0.11	0.11
agreeableness	-0.06	-0.05	0.00	0.01	-0.02
neuroticism	-0.02	-0.08	-0.05	-0.06	-0.06
social_skills	0.02	0.13	0.08	-0.05	0.06
insecure	-0.11	-0.24	-0.14	-0.10	-0.15
cold	-0.04	-0.12	-0.12	0.01	-0.08
family_oriented	-0.03	-0.07	0.02	-0.06	-0.03
adjustment	0.07	0.16	0.15	0.15	0.13
happiness	0.00	0.10	0.04	-0.04	0.04
depression	-0.04	-0.19	-0.16	-0.12	-0.11
independent	0.07	0.09	0.11	0.13	0.09
power_driven	0.09	0.20	0.06	0.10	0.11
type_a	0.00	-0.01	-0.11	0.01	-0.04
workhorse	0.08	0.17	0.05	0.08	0.12
friend_focus	-0.05	-0.09	-0.06	-0.10	-0.07
body_focus	-0.07	-0.13	-0.02	-0.13	-0.09
health_oriented	-0.04	-0.01	0.04	-0.09	0.01
sexual_focus	-0.06	-0.10	-0.07	0.00	-0.08
food_focus	-0.06	-0.17	0.04	-0.12	-0.10
leisure_oriented	-0.03	-0.03	0.05	-0.06	-0.01
money_oriented	0.11	0.20	0.14	0.10	0.10
religion_oriented	-0.02	-0.15	-0.12	-0.05	-0.07
work_oriented	0.08	0.18	0.11	0.04	0.11
netspeak	-0.02	0.00	0.05	0.03	0.01

Figure 3: Correlation between Personality and Career Progression across Industries.

For marketing and consulting industry, we have 1862 records that describe users doing marketing, advertising or consulting. The most positively correlated traits are **conscientiousness** and **extraversion**, while the most negatively correlated traits are **insecure** and **depression**. Consistent with the common knowledge, extraversion plays a more vital role (0.22) in marketing people's career, but does not affect much in IT people's career (0.04). We calculate the average score on each trait for people in different industries, it shows us that marketing people have the highest average score (30.54) on extraversion among all the groups. However, the positive correlation indicates that people who have faster career progression are less extroverted than other marketing people. Moreover, there are relatively less words describing joy or satisfaction in their posts. We can infer that people who are more introverted and less satisfied tend to move up fast in this industry.

For the media and entertainment industry, 1212 records

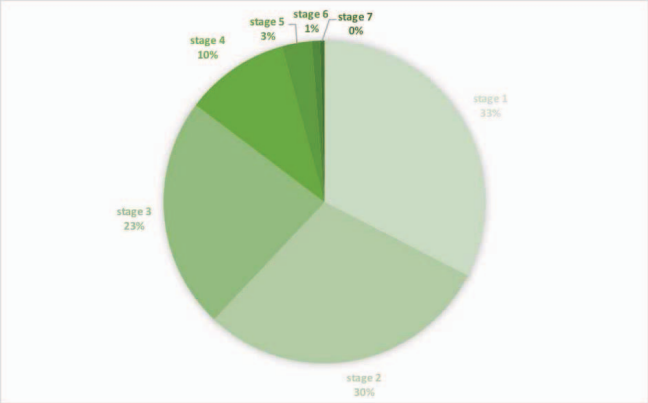

Figure 4: Distribution of people's career stages

	stage 1	stage 2	stage 3	stage 4	stage 5	stage 6	stage 7
thinking_style	0.11	0.12	0.14	0.07	0.13	0.25	0.15
persuasive	0.12	0.12	0.16	0.08	0.12	0.08	0.21
reward_bias	0.01	0.00	-0.02	0.01	0.03	-0.06	0.17
openness	-0.05	-0.07	-0.09	-0.09	-0.09	-0.32	0.10
conscientiousness	0.13	0.14	0.16	0.14	0.17	0.26	0.25
extraversion	0.11	0.09	0.10	0.05	0.12	0.25	0.30
agreeableness	-0.01	-0.01	0.01	-0.04	0.02	0.09	0.01
neuroticism	-0.06	-0.06	-0.09	-0.03	-0.11	-0.25	-0.11
social_skills	0.06	0.04	0.05	0.04	0.13	0.03	0.34
insecure	-0.14	-0.13	-0.15	-0.08	-0.17	-0.33	-0.29
cold	-0.08	-0.06	-0.07	-0.10	-0.06	0.10	-0.22
family_oriented	-0.02	-0.03	-0.04	0.02	-0.08	-0.27	-0.12
adjustment	0.12	0.13	0.15	0.09	0.16	0.24	0.18
happiness	0.04	0.01	0.01	0.03	0.07	-0.01	0.20
depression	-0.11	-0.11	-0.10	-0.03	-0.16	-0.27	-0.34
independent	0.08	0.09	0.11	0.06	0.12	0.25	0.06
power_driven	0.09	0.09	0.10	0.09	0.14	0.20	0.26
type_a	-0.05	-0.05	-0.06	-0.02	-0.08	-0.14	-0.10
workhorse	0.10	0.12	0.14	0.11	0.12	0.14	0.14
friend_focus	-0.07	-0.09	-0.08	-0.05	-0.15	-0.27	-0.02
body_focus	-0.06	-0.08	-0.12	-0.11	-0.21	-0.37	-0.09
health_oriented	-0.02	-0.01	0.00	0.05	0.05	-0.02	0.06
sexual_focus	-0.09	-0.09	-0.08	-0.05	-0.07	-0.24	0.16
food_focus	-0.08	-0.10	-0.12	-0.07	-0.14	-0.32	0.07
leisure_oriented	0.00	-0.03	-0.02	-0.03	-0.06	-0.29	0.17
money_oriented	0.08	0.08	0.07	0.09	0.10	0.31	0.19
religion_oriented	-0.07	-0.08	-0.05	-0.07	-0.13	-0.25	0.03
work_oriented	0.08	0.09	0.10	0.09	0.11	0.23	0.23
netspeak	-0.01	0.03	0.01	-0.01	0.03	0.08	0.04

Figure 5: Correlation between Personality and Career Progression across Career Stages

of promotion take place in online media, broadcast media or entertainment. The most positively correlated traits are **persuasive** and **extraversion**, and the most negatively correlated trait is **depression**. Similar to marketing people, people move up fast in media industry have higher score on depression and lower score on extraversion than others. In addition, they also make lower score on persuasive, which means their words lack intention of choosing a specific stance and persuading others. This may reflect the fact that neutrality is more welcomed and needed than extraversion and persuasion in the media industry.

For 713 records that describe users doing public relation, the most positively correlated traits are **thinking style** and **adjustment**, while the most negatively correlated trait is **body focus**. Our results demonstrate that in this industry people move up fast share relatively more informal and personal posts, their think style is more narrative but less analytical. Like people among all the industries, they lack stability in emotion and express sensibility to the environment. Moreover, body focus plays a more important role in this group of people. They tend to pay more attention to their body or even others' bodies. We may infer that public relation people's physical appearances are more critical to their jobs than people in other industries.

2) Different Career Stages: As figure 5 shows, we compute Pearson correlation coefficients to find out unique traits that impact people's promotions towards different stages. Statistically significant (p-value < 0.05) and interesting correlations are observed, and we summarize them as follows:

From Figure 4, we learn the distribution of career stages where all the promotion record take place, which matches the pattern of typical job position hierarchy. The higher the career stage, the fewer people get promoted. The statistical result further validates our method on the job level mapping. Among the eight stages, we find that generally the affects of those traits are consistent. Traits that are negatively correlated in low stages are also negatively correlated in high stages. The difference is that the traits help much or less in higher stages. For example, people who get to stage 1 faster are more insecure (-0.14), so are those who reach stage 7 faster (-0.30), while the correlation at stage 7 is much stronger. After the observation, we find that thinking style, persuasive, conscientiousness, extraversion, insecure, adjustment, depression are determinant across all the stages. However, there are particular traits that affect a lot in specific stages. For example, people who get promoted faster in stage 6 obtain higher score on openness (-0.32), family oriented (-0.27) and leisure oriented (-0.29). While these traits do not have strong correlation promotions towards career stage 1 (p-value > 0.1). Therefore, the career stage determines how much a particular trait affect the promotion prospect.

B. Education Background

It is worth mentioning that education background is also a factor that might affect career performance. Many people are going back to school to continue their education for the career advancement. Therefore, we investigate the impact of people's education background on their career paths. Figure 6 shows the histogram of average duration that people with different school ranks wait for a promotion. We again compute the Pearson correlation coefficient (-0.041) with p-value (0.014), which indicates that generally there exists linear relationship between people's school ranks and their career progression. The better the school, the slower one gets promoted. If we look into specific industries, we find negative correlation in Media (-0.099) and other smaller industries (-0.047). However, no strong linear relationship is found in popular industries like IT (-0.036), Marketing

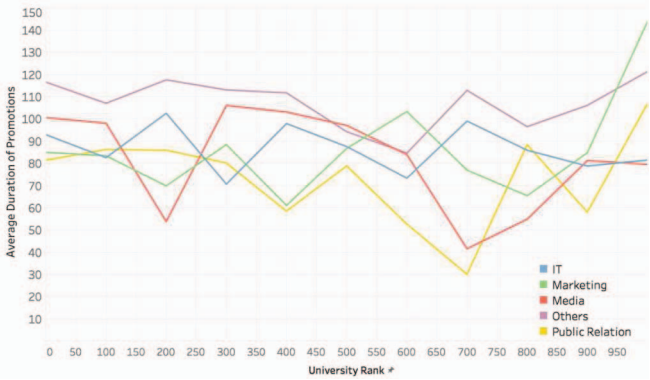

Figure 6: Average career progression duration for people of different university rankings.

Table III: Training sample industry distribution.

Information Tech	Marketing	Media	PR	Others
2008	1862	1212	713	4276

(0.005) or Public Relation (-0.031). Therefore, although better education background might help a lot when people are looking for jobs initially, it does not help much in the long career path.

V. Career Progression Prediction

Given the findings, we are inspired to predict career progression by personal traits and industry. The major learning algorithm we adopt is SVR and ensemble learning. We explore several ways to forecast the progression.

First of all, we trained our model without conducting feature (personality) selection, and we treat this as our baseline method. In this case, each training sample is described by personality vector in the form of P_1. As shown in section IV-A, people's careers in different industries are influenced by different personalities; thus, we separate our training samples by industry, and construct personality vectors by strongly correlated personalities for each industry. Table III presents the distribution of the training samples from each industry.

The personality vectors for samples from Information Technology, Marketing, Media, and PR industries are described by P_2 to P_5 respectively. The metrics we use for measuring the model performance is symmetric mean percentage error (sMAPE)[8]. We also tune our models by performing Bayesian Optimization to locate the best combination of hyper-parameters.

A. Baseline: all personalities

In the baseline case, we take all personalities into consideration. We have in total 10071 training samples, due to this size, we select ensemble tree learners as the predictor. We let X be the input matrix:

$$X = \{x^1, x^2, ..., x^{10071}\}$$

where $x^i = \{P_{1i}, industry, stage\}$, and Y be the output:

$$Y = \{y^1, y^2, ..., y^{10071}\}$$

where $y^i = \{duration_i\}$. With respect to tune the predictor, the optimization process aims to locate the combination of 1) ensemble-aggregation method, 2) number of ensemble learning cycles, 3) learning rate for shrinkage, and 4) minimal number of leaf node observations. Within 30 iterations, the best combination of hyper-parameters located along with the sMAPE is shown in Table IV, and as shown, the accuracy for our baseline method is 61.85%.

Table IV: Hyper-parameters and 10-fold cross validated sMAPE for ensemble model trained by considering all personalities. Note that learning rate does not apply to ensemble-aggregation method *Bag*.

Method	NumLearningCycles	LearnRate	MinLeafSize	sMAPE
Bag	491	NaN	2	0.3815

B. Improvement: industry-specific personalities.

Based on the discoveries from section IV-A, we separated training samples from four most popular industries: *Information Technology*, *Marketing*, *Media*, and *Public Relation*. Subsequently, we divide the training samples into five sets and construct the personality vector accordingly. Note that the fifth set includes all samples in other less popular industries, we call the set *Others*. For these four popular industries, we only select the correlated (p-value < 0.05) personalities to form P_2 to P_5.

We define the input $X^{'}$:

$$X_i^{'} = \{x^1, x^2, ..., x^j, ..., x^m\}, 1 \leq i \leq 5$$

where $m = |set_i|$ and $x^j = \{P_{ij}, stage\}$, then get the output Y':

$$Y_i^{'} = \{y^1, y^2, ..., y^j, ..., y^m\}, 1 \leq i \leq 5$$

where $y^j = \{duration_{ij}\}$. In this case, since the training size for each set become smaller, we train both SVR and ensemble tree learners for them. In the same vein, we optimize the model by Bayesian Optimization. Similar to ensemble learners, for SVR, the optimization process searches the best combination of 1) penalty, 2) kernel scale, 3) half the width of epsilon-insensitive band that produces the least 10-fold

[8] $sMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{|F_t| + |A_t|}$

Table V: Hyper-parameters tuned for best model performances and their accuracies

	SVR				Ensemble				
Industry	Box Constrain (c)	Kernel Scale	epsilon	sMAPE	Method	NumLearningCycles	LearnRate	MinLeafSize	sMAPE
IT	877.51	0.001	7172	0.396	Bag	230	-	1	0.368
Marketing	0.533	2.061	65.366	0.405	LSBoost	497	0.056	8	0.364
Media	0.001	0.001	0.286	0.391	Bag	486	-	1	0.377
PR	0.001	9.521	3565.1	0.407	Bag	301	-	2	0.373
Others	252.47	297.35	6403	0.427	Bag	494	-	2	0.390

cross validated error rate. Table V shows the best hyper-parameters found for both SVR and ensemble tree learners for each industry, plus the accuracies.

According to Table V, the ensemble learners outperform SVR in all industries. For the four most popular industries (IT, Marketing, Media, PR), the forecasting accuracies are improved by 1.35%, 1.75%, 0.45%, and 0.85%, when compared with the accuracy 61.85% baseline method, respectively. The accuracy increments are slight, but overall, considering correlated personalities by different industries demonstrates the potentials to improve our prediction goal.

VI. Limitations

Our methodology of job level mapping suits most companies and provides us promising results as described in previous sections. Although the methodology is novel and reasonable, there are a few real-life cases that are contrary to it. For example, some star startups might have very high bar than enterprises like Google. As the variation of those cases, we are not able to include them in our job level mapping system. To overcome special cases, we might consider more company factors other than the size in the future, e.g., we would collect specific job requirements including skill requirement and degree requirement from LinkedIn for job level evaluation. During the study of education background, we observe that many people do not fill out the education section on LinkedIn, and some people just leave a school name without their degrees and majors. It is difficult for us to collect integral data for analyzing other aspects of education background, such as degrees, programs and majors. We will seek other data sources to complete this task in the future. Furthermore, our prediction could be applied on people in common industries who would like to leave posts on social media, but there might be employees from traditional industries that do not frequently leave traces on social media. Nevertheless, we still cover the majority of industries and our methodology is applicable to most of the people.

VII. Conclusions and Future Work

In this paper, we discuss the impacts of personality, emotion, interests and education on people's career progression. We utilized users' information from multiple platforms. From ones' LinkedIn pages, we gather their career experience as well as education background. For each experience of a user, we also fetch the size of the company one works for. We propose a new approach to evaluating people's career stages with the consideration of job-hopping between different sized companies. Furthermore, people's tweets provided us information about their personality, interest, emotions and life attitude. We applied Receptiviti API and extracted twenty-nine traits for each user.

With the career stages well defined, we compute Pearson correlation coefficients to analyze the impact of personality, interests, education on people's career progression across different industries and career stages. The results indicate that people who move up fast in their careers share similar traits and interests. Our observation also reveals that for each industry and each stage there are unique traits help people move up faster. Finally, we employ machine learning models such as SVR, ensemble learning to predict people's next promotion based on the features extracted from tweets. A promising accuracy indicates the strong discrimination of personality features on career progression prediction tasks.

In the future, we might investigate other information of education background beyond school ranks, such as degrees and majors. We are also interested in extracting more human characteristics from tweets timeline like users' rest/sleep habits. We also would like to extract users' emotion from analyzing images they post, which might help describe a person's emotional status.

Acknowledgment

We would like to thank Professor Bo Luo of the University of Kansas for his assistance on data collection, and the support of New York State through the Goergen Institute for Data Science.

References

[1] T. L. Lindquist, L. J. Beilin, and M. W. Knuiman, "Influence of lifestyle, coping, and job stress on blood pressure in men and women," *Hypertension*, vol. 29, no. 1, pp. 1–7, 1997.

[2] K. W. Strully, "Job loss and health in the us labor market," *Demography*, vol. 46, no. 2, pp. 221–246, 2009.

[3] T. Hu, H. Xiao, J. Luo, and T.-v. T. Nguyen, "What the language you tweet says about your occupation." in *ICWSM*, 2016, pp. 181–190.

[4] S. Y. Lee, S. J. Kim, J. Shin, K.-T. Han, and E.-C. Park, "The impact of job status on quality of life: general population versus long-term cancer survivors," *Psycho-Oncology*, vol. 24, no. 11, pp. 1552–1559, 2015.

[5] N. Payne, F. Jones, and P. Harris, "The impact of working life on health behavior: the effect of job strain on the cognitive predictors of exercise." *Journal of occupational health psychology*, vol. 7, no. 4, p. 342, 2002.

[6] M. Voss, B. Floderus, and F. Diderichsen, "How do job characteristics, family situation, domestic work, and lifestyle factors relate to sickness absence? a study based on sweden post," *Journal of occupational and environmental medicine*, vol. 46, no. 11, pp. 1134–1143, 2004.

[7] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[8] T. H. Silva, P. O. V. de Melo, J. M. Almeida, M. Musolesi, and A. A. Loureiro, "You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare." in *ICWSM*, 2014.

[9] L. R. Goldberg, "The structure of phenotypic personality traits." *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[10] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PloS one*, vol. 10, no. 3, p. e0115545, 2015.

[11] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through twitter content." The Association for Computational Linguistics, 2015.

[12] G. M. Hurtz and J. J. Donovan, "Personality and job performance: The big five revisited," *Journal of applied psychology*, vol. 85, no. 6, pp. 869–879, 2000.

[13] J. F. Salgado, "The five factor model of personality and job performance in the european community." 1997.

[14] T. A. Judge, D. Heller, and M. K. Mount, "Five-factor model of personality and job satisfaction: a meta-analysis." 2002.

[15] J. R. Bradley and S. Cartwright, "Social support, job stress, health, and job satisfaction among nurses in the united kingdom," *International Journal of Stress Management*, vol. 9, no. 3, pp. 163–182, 2002.

[16] A. Reichel and A. Pizam, "Job satisfaction, lifestyle and demographics of us hospitality industry workersversus others," *International Journal of Hospitality Management*, vol. 3, no. 3, pp. 123–133, 1984.

[17] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, "Fortune teller: Predicting your career path." in *AAAI*, 2016, pp. 201–207.

[18] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on twitter," *Journal of Research in Personality*, vol. 46, no. 6, pp. 710–718, 2012.

[19] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[20] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[21] J. Golbeck, "Predicting personality from social media text," *AIS Transactions on Replication Research*, vol. 2, no. 1, p. 2, 2016.

[22] R. Eastman, "Sizing up small-to-medium business (smb)," *Small Business Research*, 2010.