The Evolution of Occupational Identity in Twitter Biographies

Xingzhi Guo[1], **Dakota Handzlik**[1], **Jason J. Jones**[2], **Steven Skiena** [1]

[1] Department of Computer Science, Stony Brook University, New York
[2] Department of Sociology, Stony Brook University, New York
{xingzguo, skiena}@cs.stonybrook.edu, {jason.j.jones, dakota.handzlik}@stonybrook.edu

Abstract

Occupational identity concerns the self-image of an individual's affinities and socioeconomic class, and directs how a person should behave in certain ways. Understanding the establishment of occupational identity is important to study work-related behaviors. However, large-scale quantitative studies of occupational identity are difficult to perform due to its indirect observable nature. But profile biographies on social media contain concise yet rich descriptions about self-identity. Analysis of these self-descriptions provides powerful insights concerning how people see themselves and how they change over time.

In this paper, we present and analyze a longitudinal corpus recording the self-authored public biographies of 51.18 million Twitter users as they evolve over six years from 2015 to 2021. In particular, we investigate the social approval (e.g., job prestige and salary) effects in how people self-disclose occupational identities, quantifying over-represented occupations as well as the occupational transitions with respect to job prestige over time. We show that self-reported jobs and job transitions are biased toward more prestigious occupations. We also present an intriguing case study about how self-reported jobs changed amid COVID-19 and the subsequent *"Great Resignation"* trend with the latest full-year data in 2022. These results demonstrate that social media biographies are a rich source of data for quantitative social science studies, allowing unobtrusive observation of the intersections and transitions obtained in online self-presentation.

Introduction

Self-identity is a concept of *who we think we are* that directs us on how "to behave in a certain way" (Vignoles et al. 2008). The formation of self-identity is rooted in people's autobiography memory, often cultivated from personal life experiences. Occupational work is arguably the life domain to which we devote the greatest part of our time awake, thus, it is an essential part of our experience, embedding itself into our motivation, life goals and personal sense of identity. *"One life-long objective is work per se"*; as Butler (1992) suggested "Every man's work .. is always a portrait of himself."; However, not every person will develop a primary occupational identity (*e.g., "I think of myself as a mother, rather than a bank clerk."*). Likewise, not every job ultimately leads to the establishment of occupational identity (*e.g., "I do not define myself as a deliverer, although I do it for living."*).

In this paper, we consider the question of *what factors contribute to the formation and transition of occupational identity?* However, research into personal identity is difficult because self-identity is not directly observable. Although an individual's employment status is explicitly stated in tax records and other datasets, an individual's occupational identity is an internal and invisible presentation of *"who I am"*, and generally not publicly available.

But recent research (Rogers and Jones 2021; Handzlik, Jones, and Skiena 2024) have proposed a new methodology, using Twitter and other social media biographies as the proxy for self-identity. The concise (on Twitter, at most 160 characters) personal biography is a feature of many social media platforms, and provides an outlet for users to report their current self-identity. On Twitter, this biography string is part of every user's public profile, and research has explored the various and changing aspects of identity that users elect to associate themselves with (Semertzidis, Pitoura, and Tsaparas 2013; Rogers and Jones 2021).

Although users differ in what they disclose, many social media biographies express relational identity (*e.g. father/mother of, alumni of*), demographics, occupations and political affiliations. Semertzidis, Pitoura, and Tsaparas (2013) further categorized biography content in terms of average word fraction: Occupation (7.5%), Interests/Preferences/Hobbies(4.6%), Personal Info (2.6%), etc. For example, the biography:

> 25. black man. student. Films. "We are guardians of the image & that's how I see our role as DP's"

contains information about age (*25-year old*), occupation (*student*), ethnicity (*black*) and gender (*man*).

Longitudinal changes in a subject's self-reported biography allows analysis of changing self-conception over the full life course. For example, the aforementioned biography is likely to updated to reflect age and career path (e.g., *25→ 26*, *student→ cameraman*). This would record an occupational identity change from *student* to *cameraman* upon graduation. Our dataset records 435.7 million twitter biography edits, and extracts more than 287,000 job transitions between

Title A	Title B	# (A↔B)
co-founder	founder	10149
ceo	founder	8577
director	executive director	2351
assistant professor	associate professor	2154
graphic designer	freelance graphic designer	1975
software engineer	senior software engineer	1043
student	nursing student	770
accountant	chartered accountant	661
student	mechanical engineer	575

Table 1: Job transition pairs observed in Twitter biographies. These transitions reflect personal career advancement (e.g., from *assistant professor* to *associate professor*), graduation (e.g., from *student* to *engineer*) or work-status change (e.g, from *designer* to *freelance designer*). We investigate why people self-present with such occupational identities.

over 25,000 distinct job titles. We present examples of the frequent job transitions in Table 1.

Our large-scale dataset of public biographies gives researchers a new opportunity to study the self-presented identities and their evolution among Twitter users. Previous studies involve much smaller datasets, and lack the longitudinal element enabling us to track changes in the self-identity of individuals over time. Pathak, Madani, and Joseph (2021) extracted a set of words from biographies as personal identifiers for future social identity studies. Other studies (Sloan et al. 2015) use automatic tools to profile Twitter users, and observe that jobs in the creative sector (e.g., writer, artist) are over-represented in UK Twitter users. However, they only report population biases in Twitter users (e.g., over-represented occupations), but do not further investigate the reasons behind them. Meanwhile, these studies generally analyze the population at one fixed time point, and ignore the evolution in each individual over time.

In our study, we collect and analyze a multi million-subject scale temporal Twitter biographies dataset, and systematically explore self-authored occupational identities. We present our quantifiable findings to better understand how people consider occupations as a part of self-identity. To further clarify our motivation, this paper focuses on the self-disclosed *occupational identity*, instead of the mere employment status of the Twitter population.

The primary contributions of this paper include:

- *A Temporal Data Set of Self-reported Job Titles* – We have assembled a longitudinal corpus of self-reported biographies as they evolve annually over the period from 2015-2021, with 1.35 million distinct users whose biographies are written in English. From this dataset, we identified the job title changes and released an occupational transition graph with 25,033 unique jobs and 287,425 unique edges. Our dataset provides a unique resource for studying changes in self-identity, and is publicly available at https://bio-job-graph.github.io.
- *Identifying the Reasons Behind Occupational Identity* – With this rich dataset, we present a large-scale case study of occupational identities among Twitter users. Further, we quantify which occupations are over-represented in self-biographies relative to the number of people holding such positions. We show that 74.04% of over-represented jobs are above the overall median occupational prestige score, versus only 39.48% for under-represented jobs. We demonstrate that over-representation is highly correlated with job prestige, with a Spearman correlation of $\rho = 0.4858$. The correlation of over-representation with prestige is stronger than that with income, where $\rho = 0.3961$.
- *The Evolution of Identity* – We present a temporal analysis explaining how and why self-descriptions change, distinguishing life events (e.g. job promotion, graduation, etc) from lateral moves. In our study of occupation identity transitions, we discover that there is an inherent directionality towards increased occupational prestige with most transitions (e.g., from *assistant professor* to *associate professor*). Moreover, we also present an intriguing case study, reflecting changes in observed occupational identity amid COVID-19 and its aftermath.

Background and Related Work

Self-identity is a concept of *who we are* and directs us how to "behave in a certain way" (Vignoles et al. 2008), fostered in our autobiographical memory (Bluck et al. 2005). Self-identity can be realized as a knowledge representation (Kihlstrom and Klein 1994) of past events with personal experience, often characterized as a personal life story. However, due to the nature of its unobservability in practice, it is notoriously challenging to conduct quantitative research of self-identity on a large scale. Recently, (Rogers and Jones 2021; Jones 2021) proposed a new methodology to discover quantifiable insights of political identities from millions of Twitter biographies, leveraging them as a directly observable source of self-identities. Such research methodology makes it possible to study self-identity in a large-scale and computational manner.

In this section, we describe self-identity related theories, methods, and insights from the existing research of psychology and computational social science.

Self-identity in theory: Generally speaking, there are two kinds of self-identity: 1) *Personal identity* describes how an individual defines him/herself as a unique human being; 2) *Social identity* represents how an individual feels belonging to certain groups as a member. Occupational identity lies somewhere in between (e.g., a unique self belonging to a work group). (Christiansen 1999) first connected occupation to identity, suggesting that occupation contributes to identity shaping (e.g., *I am a sociologist*), and occupation provides a context for creating a meaningful life, which promotes well-being and life satisfaction. To understand work-related identity, the current studies (Ashforth, Harrison, and Corley 2008) aim to answer these key questions: 1). *What is work-related identity?*; 2). *Why does identity matters?*; 3). *How does identity evolve?* To address such questions, Kielhofner (2002) proposed the term *"Occupational Identity"* and defined it as "a composite sense of who one is and wishes to become as an occupational being".

Establishment of occupational identity: Kielhofner (Kielhofner 2002) further argued that the occupations that are satisfying and recognized within one's environment are more likely to become meaningful and central to one's life, sustaining to self-identity. According to social identification theory (Ellemers, De Gilder, and Haslam 2004; Van Dick 2004; Jackson 2002), a healthy self and social identity tend to thrive toward positive self-evaluation and link to a positive social group identity. As emphasized in occupational identity research, social approval is also key to identity establishment. For example, key occupational identity theorists suggest that (Christiansen 2004; Phelan and Kinsella 2009) "identities are fostered when individuals perceive that their chosen occupations win approval from the greater society". One quantitative measure of social approval is perceived occupation prestige (Treiman 2013; Hodge, Siegel, and Rossi 1964; Hout, Smith, and Marsden 2015), derived from job-specific factors like income and education levels. Specifically, Hout, Smith, and Marsden (2015) released occupational prestige scores based on the interviews with 1,001 individuals in 2012. These ratings range from 0 (bottom) to 100 (top), covering 860 occupational titles or 539 job categories in the 2010 Standard Occupational Classification (SOC) (Elias, Birch et al. 2010). This unique resource enables us to quantitatively measure the correlation between the occupations in identity and social approval.

Self-identity in Twitter biographies: Social media users often present their self-identities in bios. Semertzidis, Pitoura, and Tsaparas (2013) investigated how people express themselves in Twitter biographies and discovered that the most common content in biographies is occupations. Pathak, Madani, and Joseph (2021) extracted several clusters of personal identifiers from Twitter biographies by matching text patterns and suggested it as a resource for future social identity studies. Similarly, Sloan et al. (2015) extracted job titles within the bios of UK Twitter users, and found that the jobs in the creative sector (e.g., artist, writer) are over-represented. They explained this observation by the fact that Twitter is used by people who work in the creative industries as a promotional tool, and perhaps as a technical artifact from misclassifying hobbies and occupations (e.g., leisure or professional writer) . Other researchers (Mislove et al. 2011) studied the demographic distribution of the Twitter population by analyzing self-reported user names and biographies, enabling large-scale social studies involving demographic characteristics. Note that self-identity researchers focus on how people self-present themselves, rather than working to improve the accuracy of user profiling algorithms (Preoţiuc-Pietro, Lampos, and Aletras 2015; Pan et al. 2019).

Self-identity transitions over time: Personal biographies are dynamic, reflecting the life or mind changes in a person over time. Shima, Yoshida, and Umemura (2017) analyzed the timing and reasons for Twitter bio changes and found that English users most frequently make changes on their birthdays. Rogers and Jones (2021) discovered that Americans defined themselves more politically by integrating more political words in their Twitter bios over recent years. In our temporal analysis, we examine the reasons behind these modifications and discover interesting bio transitions for graduation, anniversaries and job transitions. To our best knowledge, this is also the largest resource available for studying dynamics in occupational identities.

Biography Dataset Collection and Pre-processing

We conducted a study on longitudinal occupational identity by extracting self-reported job titles from the biographies of Twitter users over six years, and further mapping the job titles to Standard Occupational Classification (SOC) [1] for specific job category information (e.g., average salary, job prestige scores, etc). We provide a thorough description of our data processing procedures below.

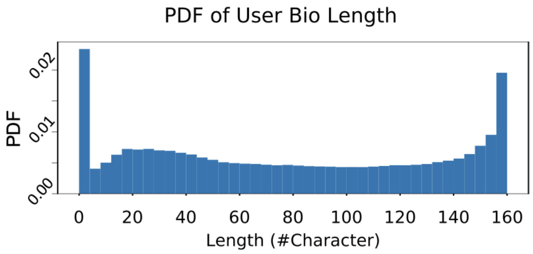

Figure 1: The statistics of sampled biography length. The two peaks indicate that people tend to either skip or write complete bios. More than ∼80% users have bios longer than 25 characters.

Collecting longitudinal Twitter biographies: We collect the Twitter user biography associated with each of 51.18 million English-language users [2] from Feb. 2015 to July 2021 using Twitter API. We have been pulling 1% of all tweets from the Twitter stream on a daily basis over this period, and have been recording personal biographies with each tweet's observed timestamp. Although our raw data is over-sampled toward more active users, we employ a de-duplication mechanism and ensure that identical user bios are counted only once, regardless of how many times they appear in our samples. In addition, our dataset, given its extensive volume and duration, also includes a substantial portion of less active users. Presumably, we should capture approximately 1% of all tweets, ensuring that we encompass users who have tweeted 100 times or more over our six-year data collection period or 1.3 tweets monthly. According to a Statista survey (Stacy Jo Dixon 2022), 96% of U.S. Twitter users claimed to use the platform monthly, indicating that our data has good coverage of all users and enables us to collaboratively build a robust dataset for analysis. We detect 435.70 million biography changes over this corpus. Figure 1 shows the bio length distribution, demonstrating that more than 80% of users take advantage of the biography field to report a substantial self-description.

[1] The U.S. SOC: https://www.bls.gov/soc/

[2] We use the FastText language identification tool (Joulin et al. 2016a,b) to filter out non-English bios.

Extracting occupation identities from biographies: Biographies contain rich occupational information (Semertzidis, Pitoura, and Tsaparas 2013). To demonstrate the usefulness of our dataset, we extract[3] and analyze self-reported job titles on a large scale. From biographies sampled over a six year period (2015-2021), we detected 25,033 unique occupational titles from 6,028,581 Twitter users, which covers 11.78% of the whole population in our dataset. To reduce the noise introduced by the business accounts and bots (e.g., user name such as Uber-Man, Morning Brew, etc.), we have chosen to employ a robust name-based filtering approach. Specifically, we only include those users whose usernames are human-like, which can be matched to US Census data [4]. Our primary objective was to maintain a high level of precision, ensuring that the majority of included accounts are indeed human, rather than focusing on recall, which aims to include all human-like accounts, even those without typical human names. Upon close examination of a subset of our data, we did not find any substantial evidence of bots. Subjects often report multiple job titles over this period, reflecting promotion or career changes.

Mapping job titles to Standardized Occupational Codes (SOC) with job prestige scores: To enhance our understanding of job titles and derive comprehensive insights such as salary and popularity, we have organized the initially unstructured textual job titles detected Leveraging the Standardized Occupational Codes (SOC) system [5] and job prestige data (Hout, Smith, and Marsden 2015) [6], which estimates occupation prestige within SOC-indexed work categories. Firstly, we performed an exact match using the SOC(2018) Direct Match Title, an exhaustive list containing 6,593 job title examples spanning all SOC occupational categories. Subsequently, we applied a text overlap ROUGE score(Lin 2004) to perform a more flexible matching of Twitter job titles to the SOC examples, associating each job title with the SOC category having the highest matching score. For precision in job category mapping, we deemed a job title unmatched if the ROUGE-F1 score fell below 0.8. Unmatched titles were excluded from our analysis to eliminate noise. To ensure alignment with the job prestige data, which is surveyed based on the OCC(2010) format, we converted SOC(2018) to OCC(2010) using official mappings.

This process yielded 1,834 unique job titles, each associated with high-precision occupational prestige scores. We illustrate several alignment examples with their corresponding prestige scores in Table 2.

Ethical Considerations and Broader Impact

Although Twitter data is public and often falls under public data exemptions in IRB consideration, sensitive information such as user's gender, ethnicity, health, political affiliation or beliefs must be used with caution. Inference on the individual level is restricted use case per the Twitter Developer Agreement, however: [7]

> Aggregate analysis of Twitter content that does not store any personal data (for example, user IDs, usernames, and other identifiers) is permitted, ...

All of our results are presented at an aggregate level, so that no individual identifier or original content can be traced back or reconstructed from our results. All potentially identifiable examples given in our results have been artificially modified to enforce privacy protection. The IRB of our institution has affirmed that the use of public information does not constitute human subjects research, in accordance with regulations from the US Department of Health and Human Services.

Opportunities in our dataset: To our best knowledge, our dataset is the first and largest resource for both conducting temporal analysis of biographies and studying occupational identities on Twitter. The potential impact of our work becomes evident through these illustrative examples: 1): In the field of Organization Science (Lee et al. 2022; Boğan and Dedeoğlu 2020; Fuller et al. 2006), our quantitative findings offer valuable insights into how organizations perceive job identity formation. This information can guide organizations in harnessing social approval mechanisms, such as enhancing job prestige and increasing social recognition, to celebrate their employees effectively. 2): Within the domain of Computer Science (Dave et al. 2018; O*NET 2023), the job transitional graph we have constructed serves as a privacy-compliant, open data source. This data facilitates modeling occupation similarity, such as in job catalog classification tasks, eliminating the need to scrape private resumes from social networks[8], thus mitigating ethical and privacy concerns. Additionally, this dataset also supports large-scale studies (Napierala and Kvetan 2023) by categorizing unstructured job titles for in-depth analysis. 3: In Economics and policy-making (Atalay et al. 2020), our research holds relevance by complementing existing datasets, like the nationally representative longitudinal surveys-derived Career Trajectories and Occupational Transitions (CTOT) dataset (Schwartz et al. 2021). Our unique approach provides a fresh perspective on this critical topic, enriching the field of Economics and contributing to well-informed policy decisions regarding workforce dynamics.

Limitations in our dataset: 1). The difference in self-disclosure may cause the result biased towards the population who are more willing to reveal themselves. 2). The job title mapping are not perfect. Although we manually validate part of the results, how to accurately annotate users on a large scale is still a challenging problem in both computer and social science studies. 3). All real-world data (job prestige, occupation distribution, standard occupation category) is based on the U.S. Among the subset of users with a non-empty Twitter location field it was observed that over 99% of them are classified as being within the U.S. using a regex-based location parser, making the results not generalizable to other regions. 4). To protect user privacy, we can only share the full job transition graph rather than the biography text.

[3] Job title extractor: https://github.com/fluquid/find_job_titles

[4] https://github.com/rossgoodwin\american-names

[5] https://www.bls.gov/soc/2018/home.htm

[6] http://gss.norc.org/Documents/other/PRESTG10SEI10_supplement.xls

[7] More about Twitter APIs: https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

[8] e.g., LinkedIn prohibits scraping its data

Job Title (Bios)	Occupation Category (SOC)	Match Score	Mention Freq.	Occupation Prestige
ceo	Chief executives	1.00	325261	71.58
civil engineer	Civil engineers	1.00	46412	65.33
coo	Chief executives	1.00	41722	71.58
mechanical engineer	Mechanical engineers	1.00	38117	70.31
accountant	Accountants and auditors	1.00	36614	59.72
...	...	...	...	...
business service representative	Sales representatives, services, all other	0.86	1	43.18
assistant food service director	Food service managers	0.86	1	39.43
medical and health service manager	Medical and health services managers	0.80	2	64.06
waste water treatment plant operator	Water and wastewater treatment plant and system operators	0.80	1	38.87

Table 2: Twitter job title mappings and job prestige estimates, ranked by match score and frequency of mention. Note that multiple job titles may map to the same SOC category. For example, both the *ceo, coo* titles fall into the category *Chief executives*, and share the same occupation prestige score. Match score is ROUGE-based with the range $[0.0, 1.0]$, where 1.0 indicates an exact match. We assign the occupation prestige score to those titles with the match score greater than 0.8 to avoid noise from job mapping noise.

Occupational Identities Reflected Through Twitter Biographies

Work is an essential event in our daily life, which contributes to our self-identity. Sociologists and occupational scientists have proposed several theories about the establishment of occupational identity. One fundamental factor is the quest for social approval (Collin et al. 2008; Taylor et al. 2017; Kielhofner 2002) (e.g., job prestige, income). Although we have a wide variety of public labor statistics (e.g., employment by sector), Twitter biographies provide a unique and directly observable view of how users define *themselves* by occupation, displaying whether the users feel their current work reflects an important part of their *self-identity*.

After removing the potential company and bot accounts, we obtain a total of 23,267 distinct self-reported job titles. We also match the job titles to Standard Occupation Code and associated job prestige scores, as illustrated in Table 2.

In the following subsections, we first walk through a few extracted occupation examples from Twitter biographies. Then, we propose to use the *Occupation Over-Represented Ratio (κ)* statistic to describe the relative difference in the frequency of occupations in Twitter biographies compared to actual real-world presence, which captures people's sense that the occupation as a part of self-identity. Finally, we analyze the job prestige variables with respect to the establishment of occupational identity.

Over/Under-represented Occupations in Twitter Biographies

The distribution of self-mentioned occupations in Twitter biographies is significantly different from that as occurs in the U.S. population at large[9]. The most/least frequently mentioned jobs in Table 3 seem contradictory to our sense of the labor market in terms of volume and socioeconomic distribution: for example, we expect more workers than executives. Prestigious occupations appear much more likely to be included as part of occupational identity, and are thus over-represented in our data. We can quantify this observation through data on the relative prestige of jobs/titles. We present our findings below.

#	Occupation	(%)	#	Occupation	(%)
1	student	12.91	23248	restaurant busser	<0.01
2	founder	6.08	23252	track sweeper	<0.01
3	director	6.06	23257	referral clerk	<0.01
4	owner	5.37	23262	soil expert	<0.01
5	ceo	4.55	23264	grocery deliverer	<0.01
6	president	4.52	23266	culinary worker	<0.01

Table 3: The most/least popular self-reported jobs.

The salary of the self-reported occupations is significantly higher: It is known that the average income of Twitter users is higher than that of the general population (Wojcik and Hughes 2019; Blank 2017). We further find that the difference in self-reported occupations in biographies makes this gap even more significant. By matching the self-reported job titles to their incomes, Figure 2 presents the histogram of the salary of the occupations in bios, showing that most users have higher salary than the median income [10]. This finding provides a clue that a job's overall salary impacts its role in developing occupational identity.

Identifying over-represented occupations in Twitter biographies: Titles can be aspirational: being CEO of Ford Motor Company means something different than CEO of a one-person startup. One interesting observation is that more people define themselves as CEOs (#5) on Twitter than Grocery Deliverers (#23264), despite the fact that there are raw fewer CEOs in the general population. This difference can

[9] We use the employment statistics from U.S. Labor Department

[10] We use Rouge-L F1 score to softly align reported occupations to New York median income list. resulting in 21,946 job titles (out of 25,033) from 3,408,916 users associated with their occupation income. To avoid the income discrepancy in different regions, we use New York's job income list as the standard. Data source: https://catalog.data.gov/dataset/occupational-employment-statistics

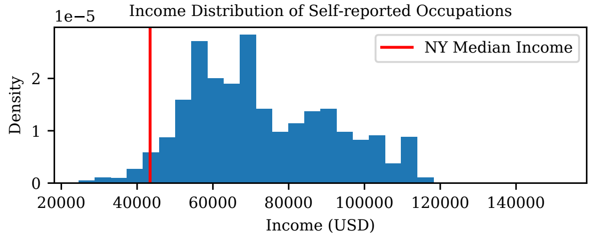

Figure 2: The salaries of self-reported jobs are significantly higher than the overall median ($43,431 USD) in New York, US. We use New York's job income statistics as standard to avoid salary discrepancy across regions.

be explained by the fact that: 1) grocery delivers may be less likely have Twitter accounts, and/or 2) grocery delivers do not consider the job as part of their self-identity and exclude it from personal biographies. In our analysis, it is the latter possibility that reveals insights about the job-specific factors toward occupational identity.

To quantify over-represented occupational families in Twitter biographies, we use the *Over-representation Ratio* κ to measure how each self-mentioned occupation exceeds the expected frequency according to the real-world workforce percentage. Specifically, κ is defined by:

$$\kappa = \frac{P(\text{Occupation}|\text{Twitter})}{P(\text{Occupation}|\text{Real-World})} \begin{cases} > 1.0 & \text{, Over-represented} \\ = 1.0 & \text{, Balanced} \\ < 1.0 & \text{, Under-represented} \end{cases}$$

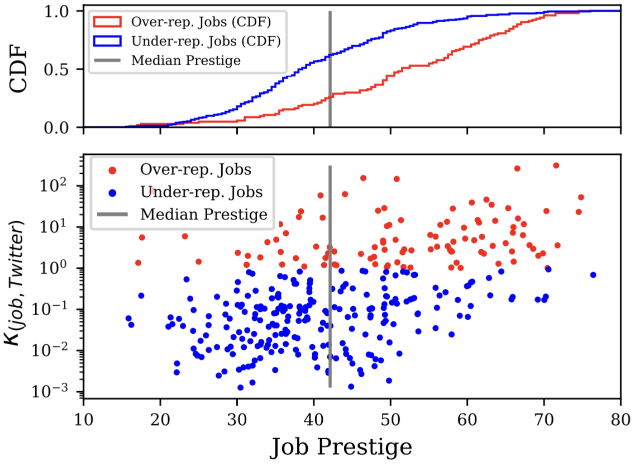

Figure 3: *Bottom:* Job prestige scores measure the socioeconomic status of an occupation according to its salary and public perception. κ measures the ratio of how an occupation is over-represented on Twitter compared to its real-world labor distribution. Prestige and κ correlate with Spearman's correlation $\rho = 0.4858$. Over-represented jobs (red) generally have higher than the median prestige. *Upper:* The prestige distribution between over/under-represented jobs are statistically different with K-S test $d = 0.4284$ and p-value $\approx 2.07 \times 10^{-12}$.

If users includes multiple occupations at at time (e.g., Professor and CEO), we handle each occupation independently and calibrate their popularity individually. To illustrate this approach: if 10% of users listed "Professor" and 20% listed "CEO," even if there is an overlap among these users, it does not impact our calibration per occupation. For instance, 10% of "Professor" on Twitter compared to just 1% in the real world [11], this indicates that the occupation "Professor" is over-represented by a factor of 10. In an extreme case, if one occupational family was never mentioned in all biographies, then $\kappa = 0.0$, indicating a completely under-represented occupational category (e.g., *track sweeper, restaurant busser, grocery deliverer*). In contrast, $\kappa \gg 1.0$ indicates a highly over-represented category on Twitter (e.g., *CEO, COO* in *Chief Executive* category).

Social Approval in Identity Establishment

The construction of occupational identity is a complex process involving personal variables and social approval. Based on the observable evidence in Twitter bios, we seek to answer the question: *Are high-prestige jobs more likely to be embedded into occupational identity?*

Prestigious jobs are more likely to be embedded into self-identity: The anecdotal evidence of Table 3, suggests that *prestigious* jobs (e.g. *CEO, COO, ...*) tend to be mentioned more frequently in personal biographies. This coincides with the hypothesis about social approval, that is, the social status of the job (including salary, and perceived prestige) is an important factor contributing to the occupational identity establishment. In Figure 3, we plot the over/under-represented groups as red/blue dots against the job prestige scores. The result demonstrates that the job prestige score is highly correlated with κ of Spearman correlation $\rho = 0.4858$. It also shows that 74.04% of over-represented jobs are above the overall median prestige score, versus 39.48% for under-represented jobs. Specifically, the prestige distributions of over/under-represented jobs are significantly different with K-S test $D = 0.4284$ and p-value $\approx 2.07 \times 10^{-12}$. Interestingly, we find that the correlation of over-representation to annual income ($\rho = 0.3961$) is weaker than it is to prestige, implying that social approval is more important than pure material return. For example, *Legislators, Clergy* and *Firefighters* share high prestige scores (with percentile > 80% in prestige), although their income percentiles are only 22.55%, 55.93% and 57.27%, respectively. However, they are all over-represented with $\kappa > 1$ based on our observations. This evidence that more prestigious jobs appear to be included more frequently in Twitter biographies as a part of one's identity.

Furthermore, we discuss the observed biases and their implications to our analysis. It is well-known that the Twitter user population is biased toward those having higher education and salary. In addition, the users from certain occupations may be more inclined to post tweets and include their jobs in bios. For instance, prestigious CEOs are over-represented in our dataset because 1): Many actual CEOs actively use Twitter as a platform for self-promotion and

[11] the numbers are for illustrative purpose

representing their companies. 2): Some individuals identify strongly with the title of CEO, even if their companies are small. Similarly, we also observed that occupations like *Tattoo Artist* may also appear over-represented, possibly due to individuals using Twitter primarily for business promotion rather than personal identity.

Our analyses capture both valid cases and compare them against real-world numbers. Although the underlying causes of over-representation are complex and multi-faceted, we are still able to extract meaningful signals because: 1: Both CEOs and Tattoo Artists can genuinely identify with their professions while utilizing Twitter for business purposes. These aspects are not mutually exclusive. 2: Our main objective in conducting large-scale data analysis is to address and mitigate the impact of noisy data. Figure 3 illustrates a clear trend, albeit not entirely free from noise. We emphasize that our motivation is **not** to uncover the real-world labor statistics from Twitter. Our contribution instead is demonstrating a clear relationship between social approval proxies and over-representation within public biographies; high-prestige and high-income jobs are over-represented while low-prestige and low-income jobs are under-represented. This evidence is consistent with the Social Approval in Identity Establishment hypothesis.

Occupational Identity Evolution

If Twitter biographies capture the notions of self-identity, then changes in biographies reflect changes in self-identify. Our dataset captures identity changes for millions of people over the six years from 2015-2021. In this section, we characterize and analyze the general biography edits, then focus quantitatively on issues of occupational identity transitions.

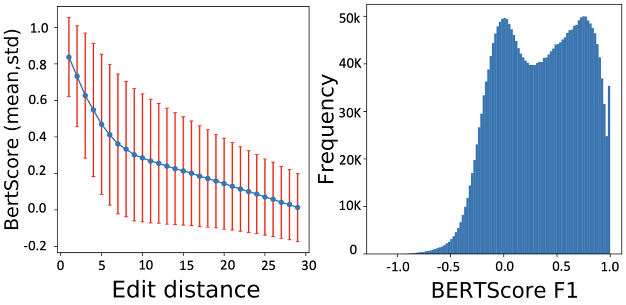

Figure 4: Left: Semantic similarity between changed bio pairs. Right: Two peaks in show that the most common changes are minor updates and complete rewrites.

We randomly sample 1,353,325 of our 51.18M individuals and acquire a subset of 2,597,550 consecutive changed pairs whose word edit distance is within 30 characters. Figure 5 shows that most people update their bios by 1-5 words at a time. But character differences do not completely reflect semantic differences: an "artist" transformed in a "painter" represents a smaller change than the edit distance suggests. To better capture semantic similarity, we extend this analysis to all changed pairs with $BERTScore_{F1} \in [-1, 1]$ ((Zhang et al. 2019)) as the semantic similarity measure.

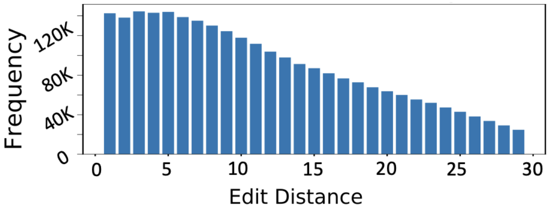

Figure 5: Small bio changes are more common than complete rewrites.

Figure 4 shows the BERTScore against word-level Levenshtein distance. As expected, semantic similarity decreases as more edits are applied. Figure 4 shows the distribution of BERTScore, where the hump centered at 0 implies that a significant fraction of people completely rewrite their biographies with each transition, while the hump at 0.8 shows another large group of people making minor edits to their bios (e.g: age or job title updates).

POS change (before, after)	Count
(NUM, NUM)	23,335
(NOUN, NOUN)	15,850
(X, X)	12,889

Table 4: The most popular change is between numbers, reflecting age updates on birthdays or anniversaries. X is not recognized, e.g.: Emoji or unrecognized abbreviation. Total number of replaced pairs is 89,838.

Minor Edits in Biographies

To understand the nature of the edits distribution, we look into the changed biography pairs where only single words have been substituted (89,838 pairs), and discover many occupational transitions over time. First, we present the top part-of-speech (POS) changes in Table 4. These changes usually share the same POS tag, indicating the alternation of word choices.

Breakdown of Nouns: To better understand the reasons behind word replacement, we organize interpretable relations between changed nouns in Table 5. We assign five categories of noun changes based on syntactic features (e.g: same lemma). Besides *Others* group, most of them are minor changes (with single character change), reflecting job promotion (e.g. *vp* to *svp*) or age updates. In addition, we observed many career transitions (e.g., from *manager* to *director*) in all categories, which is addressed in the next section.

Transitions Between Occupational Identities

To better understand the changes of occupational identities, we extract and analyze those biography pairs in which any occupational title has been added or removed.

Represent occupational transitions as a graph: We formulate occupational transitions as a directed weighted graph

Category	Total	%	Examples (Before, After, #)
Share lemma	693	4.54%	(alumni, alumnus, 60)
Synonym	669	4.38%	(manager, director,41)
One Token Change	2327	15.89%	(vp, svp, 20); (22yrs, 23yrs, 14)
Hyponymn	335	2.19%	(student, graduate, 46)
Hypernymn	266	1.17%	(teacher, educator, 17); (journalist, writer, 8)
Others	10898	71.36%	(sophomore, junior, 68); (coordinate, manager, 28)
All	15272	100%	-

Table 5: Breakdown of changed nouns. One pair may belong to more than one category. Some of them are related to the change of occupational identity (coordinate -> manager).

$G(V, E)$, where V contains all unique job titles as node, E contains all job title transitions $e_{i,j} = (v_i, v_j, w_{i,j})$, where edge weight $w_{i,j}$ represents the total number of the observed transitions from i to j. For example, a biography changed from *"Student, Volunteer"* to *"Software Engineer, Volunteer"* increases the edge weight of *Student* → *Software Engineer*. Our final graph contains 25,033 nodes and 287,425 unique edges, consisting of seven weakly-connected components. The largest component contains 25,021 nodes (99.95% of all). We illustrate a subgraph in Figure 6.

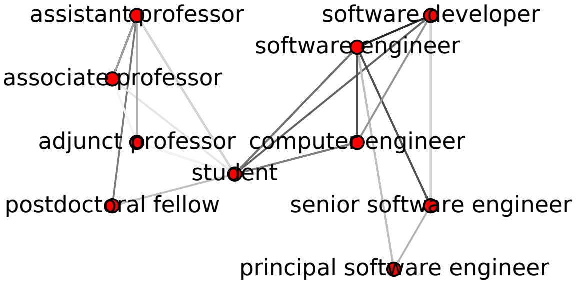

Figure 6: Examples of job transitions in the graph. Darker edges represent more frequently observed transitions.

Community of Occupational Identity Transitions: The job transition graph has an appealing property of clusterness, which reflects the proximity between occupational titles. For example, in the rooted clusters from *Student* in Figure 6, the roles of academia are naturally clustered on the left-hand side (the source), while IT industrial roles are intertwined on the right-hand side (the sink), but well-separated from the academic counterpart.

Graph partitioning is well-studied for community detection based on modularity (Brandes et al. 2007) or Personalized PageRank (Chen et al. 2023). We partition the job graph (unweighted, undirected) using agglomerative algorithm (Clauset, Newman, and Moore 2004), discovering densely intra-connected communities. Table 6 shows that the related job titles generally fall into the same cluster. Some of these share common text patterns (e.g., *"hadoop developer"* and *"hadoop architect"*), while others are inherently related but without significant text overlap (e.g., *"highway maintenance worker"* and *"bridge inspector"*). These detected transitions and clusters can provide more context to better understand personal social dynamics. The full graph can serve as a unique resource for better job title clustering, which we will release upon publication.

health information director, health information technician
highway maintenance worker, bridge inspector, toll collector, highway inspector
professor of early childhood education, peer educator, speech language pathologist assistant,
communication instructor, language specialist, public speaking teacher
customer logistics manager, import manager, export sales manager, export specialist

Table 6: Sampled job titles in graph clusters: each cluster (row) contains several highly related job titles derived from our occupational transition graph.

Most importantly, this clusterness demonstrates that mobility in occupational identity is usually limited to the same job category, partially because occupational identity directs us to behave as a member of a particular group. For example, a Professor in Computer Science rarely becomes a software engineer even though s/he may be able to do so for greater material return.

Directionality in Self-reported Job Transitions: We identify the most common title transitions (changing from Job-*i* to Job-*j* in either direction) in Table 7 (top), and find that the most frequent transitions relate to title revisions (e.g., *vice president* to *vp*). Furthermore, we define the transition *directionality*, the fraction of directed transitions from v_i to v_j, as :

$$dir_{i,j} = \frac{w_{i,j}}{w_{i,j} + w_{j,i}} \in [0, 1], \tag{1}$$

We sort the transitions by $dir_{i,j}$ in Table 7 (middle), showing that the most directed transitions are related to job promotion and reflect a significant role change (eg., *"student to nurse"*, *"postdoc to professor"*). Meanwhile, among the *less directed transitions*, most of the job titles are interchangeable with subtle difference in specificity (*scientist* to *neuroscientist*). The discovery of this directionality better quantifies the relationships between jobs, and is helpful for understanding job mobility.

Directionality Correlates with Job Prestige: One possible hypothesis is that job updates always reflect increases in job prestige. We calculate the prestige difference $\delta_{(i,j)}$ within each job transition pair, and plot against the transition directionality in Figure 7, where we observed a clear positive correlation with Pearson's $\rho = 0.3686$. This finding supports that *"we strive for positive work-related allegiances,* *enhance well-being ... and promote career development"*, as suggested by Gecas (1982) and echoed in (Bartel and Dutton 2001; Cheng, Sanchez-Burks, and Lee 2008; Ibarra 1999).

Title i	Title j	#	$dir_{i,j}$
	Top Mentioned Transitions		
co-founder	founder	10149	0.54
ceo	founder	8577	0.57
owner	founder	7384	0.59
owner	ceo	5917	0.53
director	founder	2987	0.57
	Highly Directed Transitions		
student nurse	staff nurse	109	0.99
nursing student	registered nurse	217	0.98
assistant professor	associate professor	2154	0.97
postdoctoral fellow	assistant professor	136	0.96
account manager	account director	163	0.96
	Less Directed Transitions		
director	business director	142	0.50
scientist	neuroscientist	126	0.50
assistant coach	asst. basketball coach	110	0.50
graphic designer	video editor	110	0.50
cto	owner	110	0.50

Table 7: The detected occupation transition pairs: generally the transition reflects career changes, graduation (e.g., *student* to *nurse*) or just rephrasing (e.g., *vice president* to *vp*). The directionality indicates the transition is biased toward one way, which usually indicates a positive movement (e.g., being promoted from *manager* to *director*.)

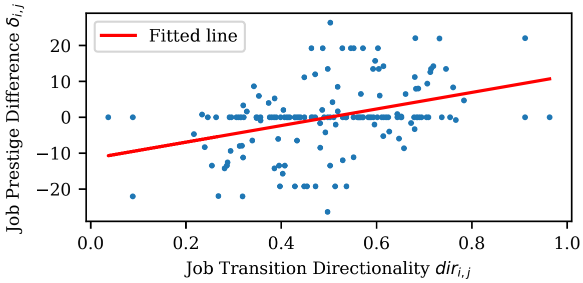

Figure 7: Highly-directed job transitions generally tend towards more prestigious positions. Some prestige scores do not change because of the limited granularity of the job prestige data (e.g., *assistant prof.* and *associate prof.* have the same prestige score).

Case Study: Transitions in/out Occupational Identities Space amid COVID-19

We performed a case study to qualify how users move in and out of occupational identity space. In order to cover the complete pandemic period, we augmented our dataset with full-year data for 2022. With regards to the graph structure, the goal was to identify which jobs are most overrepresented in (none->job) and (job->none) transitions. Table 8 shows the breakdown of these transitions by type. Among 13,980,687 unique U.S. users there were 1,602,731 (∼11.5%) who listed a recognizable job title in their bio at some point during the observation period.

Type	Counts	Proportion
job→job	847,412	0.23
job→none	1,402,957	0.38
none→job	1,446,278	0.39

Table 8: Transition Types. job->none and none->job transitions share nearly equal proportions of total transitions.

Rank	Entry Jobs	Counts	Exit Jobs	Counts
1	Student	84,489	Student	95,848
2	Owner	65,162	Owner	60,284
3	President	58,465	President	57,441
4	Founder	51,650	Founder	49,124
5	CEO	49,004	CEO	43,527

Table 9: Top entry/exit jobs by raw counts. All job titles follow the same ranking in both categories.

Table 9 shows that jobs share the highest entry/exit counts, revealing that the top occupations in each category are the same, and even follow the same ranking order. Combined with the average number of transitions per user being just over two, this indicates that a large portion of these transitions consist of "flip-flops", where a user changes their bio away from a listed occupation, perhaps for something seasonal or topical, and then returns to their original bio by the next observation. For the purposes of this analysis such flip-flops are strictly a source of noise: to compensate for this, raw counts were abandoned in favor of net changes as defined in equation 2. Overrepresented job titles were only considered if they appeared in more than 1/10000 occupational identities.

$$\delta_{net}(Job) = \#(None \rightarrow Job) - \#(Job \rightarrow None) \quad (2)$$

Personal care and service jobs are rising: Table 10 presents the most overrepresented entry/exit jobs over the full observation period. Half of the top 10 entry jobs fall under the category of personal care and service occupations, including esthetician, licensed esthetician, doula, health advocate, and tattoo artist. According to the U.S. Bureau of Labor and Statistics this group of jobs is projected to grow 14% from 2021 to 2031, nearly double the growth rate of the job market overall [12]. Other titles within this list (dominatrix, forex trader) correspond to more recent growth trends, particularly during the pandemic [13] [14].

[12] https://www.bls.gov/ooh/personal-care-and-service/home.htm

[13] https://www.nytimes.com/2021/04/10/style/findom-kink.html

[14] https://financialit.net/news/people-moves/ironfx-hires-200-employees-after-massive-rise-forex-accounts-amid-covid-19

Rank	Entry Jobs	Exit Jobs
1	Esthetician	Soccer Player
2	Licensed Esthetician	Multimedia Journalist
3	Assistant Baseball Coach	Editor-in-Chief
4	First Responder	Independent Consultant
5	Doula	Public Relations
6	Health Advocate	Snowboarder
7	Tattoo Artist	Student
8	Dominatrix	Song Writer
9	Assistant Professor	YouTuber
10	Forex Trader	Associate Professor

Table 10: Top 10 Overrepresented entry/exit jobs. Positive net change $\delta_{net}(Job) > 0$: we observed more people establishing these work identities than giving them up.

***Aspirational* identities are easier to be abandoned**: The top 10 exit jobs largely correspond to media positions (multimedia journalist, editor-in-chief), sports positions (soccer player, snowboarder), and entertainment (song writer, YouTuber). These reflect a category which can be described as "aspirational jobs"; careers with low barriers to entry and title adoption but higher barriers for conversion to a long-term professional vocation. These aspirational jobs are also congruent with the most desired jobs among both children and teens according to several surveys [15] Aspirational jobs are clearly over-represented on Twitter with respect to self-presented occupational identity. Other titles on the list, such as student, represent a position traditionally intended to be temporary. In these contexts it matches expecations to see users dropping these titles at a higher rate than more conventional long-term occupations.

Certain types of entry jobs significantly declined amid COVID-19: Figure 8 shows the changes in entry/exit job pool sizes over time. When considering unique job titles, the pool of possible entry jobs is consistently larger than the pool of possible exit jobs. Put another way, there are more unique ways to enter into occupational identity space than there are to leave it. This relationship holds true across the full observation period, but during the course of the COVID-19 pandemic the pool sizes draw significantly closer together, with the relationship nearly reversing during peak unemployment [16]. COVID-19 exacerbated trends which have not yet recovered.

Occupational Identity Rates During COVID-19:

Figure 9 shows the proportions of valid job titles among the users who exhibit at least one transition and listed an occupational identity at least once within the observation period. Users with occupational identities show a greater sensitivity to actual employment rates, an effect which disappears in the total user cohort. Although in this paper our main goal is not to conduct real-world job market research using Twitter data, these aforementioned case study results demonstrate that our proposed dataset is a reliable and useful resource for future work-related studies.

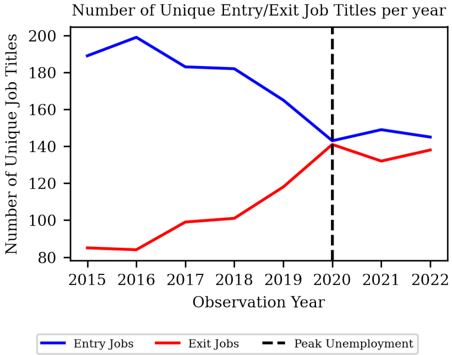

Figure 8: The number of unique entry job titles is consistently larger than the number of unique exit job titles. However, this relationship begins to falter during the COVID-19 pandemic, indicating that people were removing a wider variety of jobs from their occupational identities during the many employment disruptions surrounding the pandemic and associated lockdowns.

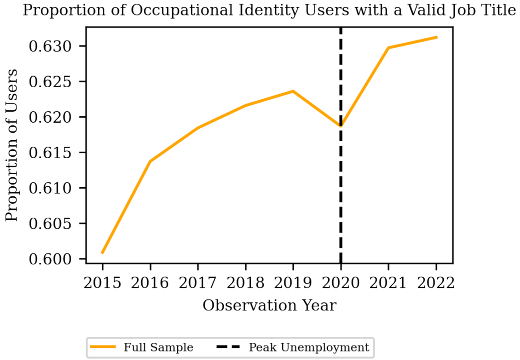

Figure 9: Among the cohort who have historically included an occupational identity there is a clear local minimum corresponding to the peak U.S. unemployment rate at the beginning of the COVID-19 pandemic. This indicates that the reported status of people who already subscribed to an occupational identity does share some correlation with actual employement status.

Conclusions and Future Work

We have presented a study of occupational identities and their evolution in Twitter biographies from 2015 to 2021, and beyond. We described a new temporal self-identity dataset, consisting of extracted job titles with associated metadata (e.g., income, prestige score), and we released the occupation transition graph. Our quantitative analysis reveals a clear correlation between high-prestige occupations and their prominence in self-identities, as well as a bias toward transitions into more prestigious roles. Additionally,

[15] https://arstechnica.com/science/2019/07/american-kids-would-much-rather-be-youtubers-than-astronauts/

[16] as reported by the U.S. Bureau of Labor and Statistics, https://www.bls.gov/charts/employment-situation/civilian-unemployment-rate.htm

we offer a compelling case study highlighting the impact of COVID-19 on work identities, showcasing the dataset's ability to capture intriguing real-world phenomena. To our knowledge, this is the most comprehensive study of social media biography dynamics to date.

We anticipate future research to build upon our findings in several key directions 1): Developing job title embeddings from the dynamic occupational graph (Guo, Zhou, and Skiena 2021) to capture job title relationships and identify novel job changes (Guo, Zhou, and Skiena 2022; Zhang et al. 2023). 2): Investigating the persistence of self-identities, with a focus on distinguishing ephemeral from enduring identities. Aligning with the efforts to understand the revisions in news headlines (Guo et al. 2022), we aim to uncover the mechanisms driving self-description revisions, in the broad context of identities. Certain identities, such as parental roles like "Father/Mother, may display greater variability in comparison to others. 3):It is intriguing to investigate job dynamics within the generative AI context, particularly concerning the emergence and sustainability of new professions like AI Artists. Our hypothesis aligns with the notion that "aspirational identities" are more likely to be abandoned over time. Additionally, we anticipate a surge in the popularity of certain roles, such as Prompt Engineer and LLM Scientist, due to increased social attention, approval, and prestige in the GenAI era.

Acknowledgments

This work was partially supported by NSF grants IIS-1926781, IIS1927227, IIS-1546113 and OAC-1919752. Happy birthday, Luca Guo, welcome to the world.

References

Ashforth, B. E.; Harrison, S. H.; and Corley, K. G. 2008. Identification in organizations: An examination of four fundamental questions. *Journal of management*, 34(3).

Atalay, E.; Phongthiengtham, P.; Sotelo, S.; and Tannenbaum, D. 2020. The evolution of work in the United States. *American Economic Journal: Applied Economics*, 12(2).

Bartel, C.; and Dutton, J. 2001. Ambiguous organizational memberships: Constructing organizational identities. *Social identity processes in organizational contexts*, 115–130.

Blank, G. 2017. The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, 35(6): 679–697.

Bluck, S.; Alea, N.; Habermas, T.; and Rubin, D. C. 2005. A tale of three functions: The self-reported uses of autobiographical memory. *Social cognition*, 23(1): 91–117.

Boğan, E.; and Dedeoğlu, B. B. 2020. Hotel employees' corporate social responsibility perception and organizational citizenship behavior: Perceived external prestige and pride in organization as serial mediators. *Corporate Social Responsibility and Environmental Management*, 27(5).

Brandes, U.; Delling, D.; Gaertler, M.; Gorke, R.; Hoefer, M.; Nikoloski, Z.; and Wagner, D. 2007. On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2): 172–188.

Butler, S. 1992. *The way of all flesh*. 118. Everyman's Library.

Chen, Z.; Guo, X.; Zhou, B.; Yang, D.; and Skiena, S. 2023. Accelerating Personalized PageRank Vector Computation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 262–273.

Cheng, C.-Y.; Sanchez-Burks, J.; and Lee, F. 2008. Connecting the dots within: Creative performance and identity integration. *Psychological Science*, 19(11): 1178–1184.

Christiansen, C. 2004. Occupation and identity: Becoming who we are through what we do. *Introduction to occupation: The art and science of living*, 121–139.

Christiansen, C. H. 1999. Defining lives: Occupation as identity: An essay on competence, coherence, and the creation of meaning. *The American Journal of Occupational Therapy*, 53(6): 547–558.

Clauset, A.; Newman, M. E.; and Moore, C. 2004. Finding community structure in very large networks. *Physical review E*, 70(6): 066111.

Collin, K.; Paloniemi, S.; Virtanen, A.; and Eteläpelto, A. 2008. Constraints and challenges on learning and construction of identities at work. *Vocations and Learning*, 1(3).

Dave, V. S.; Zhang, B.; Al Hasan, M.; AlJadda, K.; and Korayem, M. 2018. A combined representation learning approach for better job and skill recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Elias, P.; Birch, M.; et al. 2010. SOC2010: revision of the Standard Occupational Classification. *Economic & Labour Market Review*, 4(7): 48–55.

Ellemers, N.; De Gilder, D.; and Haslam, S. A. 2004. Motivating individuals and groups at work: A social identity perspective on leadership and group performance. *Academy of Management review*, 29(3): 459–478.

Fuller, J. B.; Hester, K.; Barnett, T.; Frey, L.; Relyea, C.; and Beu, D. 2006. Perceived external prestige and internal respect: New insights into the organizational identification process. *Human relations*, 59(6): 815–846.

Gecas, V. 1982. The self-concept. *Annual review of sociology*, 1–33.

Guo, X.; Kondracki, B.; Nikiforakis, N.; and Skiena, S. 2022. Verba volant, scripta volant: Understanding post-publication title changes in news outlets. In *Proceedings of the ACM Web Conference 2022*, 588–598.

Guo, X.; Zhou, B.; and Skiena, S. 2021. Subset node representation learning over large dynamic graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 516–526.

Guo, X.; Zhou, B.; and Skiena, S. 2022. Subset node anomaly tracking over large dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 475–485.

Handzlik, D.; Jones, J. J.; and Skiena, S. 2024. HINENI: Human Identity across the Nations of the Earth, Ngram Investigator. In *International AAAI Conference on Web and Social Media*.

Hodge, R. W.; Siegel, P. M.; and Rossi, P. H. 1964. Occupational prestige in the United States, 1925-63. *American Journal of Sociology*, 70(3): 286–302.

Hout, M.; Smith, T. W.; and Marsden, P. V. 2015. Prestige and socioeconomic scores for the 2010 Census codes. *Methodological Report MR124, Chicago, NORC. http://gss. norc. org/get-documentation/methodological-reports*.

Ibarra, H. 1999. Provisional selves: Experimenting with image and identity in professional adaptation. *Administrative science quarterly*, 44(4): 764–791.

Jackson, J. W. 2002. Intergroup attitudes as a function of different dimensions of group identification and perceived intergroup conflict. *Self and identity*, 1(1): 11–33.

Jones, J. J. 2021. A dataset for the study of identity at scale: Annual Prevalence of American Twitter Users with specified Token in their Profile Bio 2015–2020. *PloS one*, 16(11).

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016b. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Kielhofner, G. 2002. *A model of human occupation: Theory and application*. Lippincott Williams & Wilkins.

Kihlstrom, J. F.; and Klein, S. B. 1994. The self as a knowledge structure. *Handbook of social cognition*, 1: 153–208.

Lee, B. Y.; Kim, T.-Y.; Kim, S.; Liu, Z.; and Wang, Y. 2022. Socially responsible human resource management and employee performance: The roles of perceived external prestige and employee human resource attributions. *Human Resource Management Journal*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Napierala, J.; and Kvetan, V. 2023. Changing job skills in a changing world. In *Handbook of Computational Social Science for Policy*. Springer International Publishing Cham.

O*NET. 2023. O*NET OnLine.

Pan, J.; Bhardwaj, R.; Lu, W.; Chieu, H. L.; Pan, X.; and Puay, N. Y. 2019. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Pathak, A.; Madani, N.; and Joseph, K. 2021. A Method to Analyze Multiple Social Identities in Twitter Bios. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Phelan, S.; and Kinsella, E. A. 2009. Occupational identity: Engaging socio-cultural perspectives. *Journal of Occupational Science*, 16(2): 85–91.

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1754–1764.

Rogers, N.; and Jones, J. J. 2021. Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? *Journal of Social Computing*, 2(1): 1–13.

Schwartz, D.; Clarkwest, A.; Hashizume, M.; Kappil, T.; and Strawn, J. 2021. Building Better Pathways: An Analysis of Career Trajectories and Occupational Transitions.

Semertzidis, K.; Pitoura, E.; and Tsaparas, P. 2013. How people describe themselves on Twitter. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*, 25–30.

Shima, J.; Yoshida, M.; and Umemura, K. 2017. When do users change their profile information on twitter? In *2017 IEEE International Conference on Big Data (Big Data)*, 3119–3122. IEEE.

Sloan, L.; Morgan, J.; Burnap, P.; and Williams, M. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, 10(3).

Stacy Jo Dixon. 2022. Frequency of Twitter use in the United States as of 3rd quarter 2020.

Taylor, R.; Pan, A.; Kielhofner, G.; and Taylor, R. 2017. Doing and becoming: Occupational change and development. *Kielhofner's model of human occupation*, 140–156.

Treiman, D. J. 2013. *Occupational prestige in comparative perspective*. Elsevier.

Van Dick, R. 2004. My job is my castle: Identification in organizational contexts. *International review of industrial and organizational psychology*, 19: 171–204.

Vignoles, V. L.; Manzi, C.; Regalia, C.; Jemmolo, S.; and Scabini, E. 2008. Identity motives underlying desired and feared possible future selves. *Journal of personality*, 76(5).

Wojcik, S.; and Hughes, A. 2019. Sizing up Twitter users. *PEW research center*, 24.

Zhang, C.; Xiang, W.; Guo, X.; Zhou, B.; and Yang, D. 2023. SubAnom: Efficient Subgraph Anomaly Detection Framework over Dynamic Graphs. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1178–1185. IEEE.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Paper Checklist

1. For most authors...
   - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   - (e) Did you describe the limitations of your work? Yes
   - (f) Did you discuss any potential negative societal impacts of your work? N/A
   - (g) Did you discuss any potential misuse of your work? Yes
   - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   - (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   - (b) Have you provided justifications for all theoretical results? Yes
   - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   - (e) Did you address potential biases or limitations in your theoretical framework? Yes
   - (f) Have you related your theoretical results to the existing literature in social science? Yes
   - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   - (a) Did you state the full set of assumptions of all theoretical results? N/A
   - (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   - (a) If your work uses existing assets, did you cite the creators? Yes
   - (b) Did you mention the license of the assets? No, because they are common tools in public libraries
   - (c) Did you include any new assets in the supplemental material or as a URL? Yes
   - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes
   - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset ? No, but it can be in our plan when we release the data