Original Article

Sociological Methods & Research
1–36
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241231168665
journals.sagepub.com/home/smr
SAGE

A Method for Estimating Individual Socioeconomic Status of Twitter Users

Yuanmo He
and Milena Tsvetkova

Abstract

The rise of social media has opened countless opportunities to explore social science questions with new data and methods. However, research on socioeconomic inequality remains constrained by limited individual-level socioeconomic status (SES) measures in digital trace data. Following Bourdieu, we argue that the commercial and entertainment accounts Twitter users follow reflect their economic and cultural capital. Adapting a political science method for inferring political ideology, we use correspondence analysis to estimate the SES of 3,482,652 Twitter users who follow the accounts of 339 brands in the United States. We validate our estimates with data from the Facebook Marketing application programming interface, self-reported job titles on users' Twitter profiles, and a small survey sample. The results show reasonable correlations with the standard proxies for SES, alongside much weaker or nonsignificant correlations with other demographic variables. The proposed method opens new opportunities for innovative social research on inequality on Twitter and similar online platforms.

Department of Methodology, The London School of Economics and Political Science, London, UK

Corresponding Authors:
Yuanmo He, Department of Methodology, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.
Email: y.he54@lse.ac.uk

Milena Tsvetkova, Department of Methodology, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.
Email: m.tsvetkova@lse.ac.uk

Keywords

socioeconomic status, Twitter, correspondence analysis, measurement, social media, cultural capital

Introduction

Socioeconomic status (SES), a concept that describes people's social and economic position relative to others, is one of the most fundamental concepts in social science, underlying major areas of research such as health, education, psychology, sociology, and public policy (Diemer et al. 2013; Krieger, Williams, and Moss 1997; Oakes and Andrade 2017; Rodríguez-Hernández, Cascallar, and Kyndt 2020). Some researchers focus on measures of SES, in an attempt to capture the social stratification of modern society (Chan and Goldthorpe 2007a; Hauser and Warren 1997; Savage et al. 2013), while others investigate how SES relates to other life outcomes and thus propagates socioeconomic inequality. We know, for example, that people's SES affects their physical and mental health (Adler et al. 1994; Dohrenwend et al. 1992), political participation (Brady, Verba, and Schlozman 1995; Milligan, Moretti, and Oreopoulos 2004), the size and diversity of their social circle (Campbell, Marsden, and Hurlbert 1986; Marsden 1987), and their access and use of information and communication technologies (van Deursen and van Dijk 2014; van Deursen and Helsper 2015; Hargittai and Hinnant 2008). Most notably, people's SES is highly predictive of their children's SES, outlining the major pathway through which inequality is transmitted, social mobility constrained, and advantage accumulated across generations (DiPrete and Eirich 2006; Sirin 2005).

Most of the existing quantitative research on SES and socioeconomic inequality relies on statistical models of survey, census, and administrative-record data. The recent rise of computational social science (CSS), however, offers opportunities to study socioeconomic inequality with an entirely different set of tools and data—applying text analysis, network analysis, or machine learning methods to web, mobile, or satellite "digital trace" data (Lazer et al. 2009). For example, CSS researchers have combined night-time maps with high-resolution daytime satellite images to estimate poverty in regions with poor administrative data (Abitbol and Karsai 2020; Jean et al. 2016). Scientists have also analyzed aggregate data on Google searches and daily usage patterns of Twitter to predict unemployment claims before official statistics are released (Choi and Varian 2012; Llorente et al. 2015). Others have used social media and mobile network data to link economic development to social capital, showing that individuals who live in areas with a high local development index tend to have more diverse networks (Eagle, Macy, and Claxton 2010) with bridges that span greater geographic distances (Norbutas and Corten 2018).

These CSS studies on socioeconomic inequality, however, are conducted at the level of geographic units. Large-scale individual-level analyses using digital trace data are less common since researchers rarely have access to users' demographic and financial information. One notable exception is a unique dataset that couples mobile phone communication with bank transaction history for a subsample of the population of a Latin American country (Leo et al. 2016, 2018; Luo et al. 2017). Another prominent exception is a recent research collaboration between high-profile social scientists and Facebook, granting access to rich individual information for millions of the online social network's US users (Chetty et al. 2022). Data like these, however, tend to be proprietary and not easily accessible.

To address this gap, computer scientists have developed various methods for inferring demographic attributes from openly available digital-trace data; however, very few of these concern SES, social class, and their indicators: income, education, and occupation (Hinds and Joinson 2018). Researchers are yet to find an effective, theoretically grounded, and scalable method to infer the individual-level SES of online users. Such a method will allow linking measures of SES to the detailed records of everyday decisions, behaviors, opinions, and interactions that digital-trace data offer. The resulting research will provide population-level natural-setting observations of the daily reproduction of socioeconomic inequality. A better understanding of how limited financial resources and education may drive self-defeating behavior, strain interactions with others, or restrict access to valuable information will empower us to tackle inequality from the bottom up, complementing top-down legislative and policy reforms.

The current paper addresses the identified gap in the literature by outlining a method to estimate the SES of individual Twitter users. Twitter is a social media platform with 1.3 billion accounts and 330 million monthly active users, where 500 million tweets are posted per day (Brandwatch 2020). It is one of the most popular social media platforms used for CSS research: the number of Twitter-related studies is consistently growing (see reviews by Karami et al. 2020; McCormick et al. 2017; Yu and Muñoz-Justicia 2020). The public messaging aspect of Twitter provides valuable opportunities for researchers to observe behaviors, social interactions, and networks with a minimum obtrusion, in real time, at a low cost, and on a large scale. Moreover, Twitter offers a well-developed application programming interface (API) that makes the data more accessible compared to other popular digital platforms (e.g., Facebook, Instagram, and TikTok).

Nevertheless, it is hard to infer Twitter users' SES. Twitter does not have a designated field that requires socioeconomic information. Some Twitter users state their occupation in their profile description field, but few disclose this information accurately or at all (Sloan et al. 2015). Reviews on the topic show that existing studies on estimating the SES of individual Twitter users are scarce and disparate, and most of them have methodological limitations (Ghazouani et al. 2019; Hinds and Joinson 2018).

In this paper, we present a method to estimate the SES of individual Twitter users from the commercial and entertainment accounts they follow on the platform. The method parallels an established political science approach that uses correspondence analysis (CA) to estimate Twitter users' political ideology from the politicians and news media they follow (Barberá 2015; Barberá et al. 2015). In accordance with Bourdieu's (1984) multidimensional definition of social class, the proposed measure of SES aims to capture a combination of economic and cultural capital. As economic and cultural practices may differ in different countries, we here present the method using popular brands in the US and US Twitter users only. With the information from the Twitter accounts of 339 brands and their followers, we are able to estimate the SES of 3,482,652 users. We validate our estimation with brand consumer statistics from Facebook, self-described occupation from thousands of Twitter profiles, and survey responses on education and income from a small sample of Twitter users. Although further fine-tuning and external validation will be desirable, our preliminary results indicate that the method promises to become a valid and useful measure of SES for Twitter users.

Measuring SES: From Survey Data to Twitter

The idea to approach modern societies as strata or segments of SES groups is one of the most fundamental and deeply rooted ideas in sociology, tracing its origins back to Durkheim, Marx, and Weber. Yet, 150 years later, the problem of how to define and measure SES is still contested and unresolved. There are debates regarding whether SES is unidimensional or multidimensional and what to include in the measure (Chan 2019a; Chan and Goldthorpe 2007a; Flemmen, Jarness, and Rosenlund 2019; Hauser and Warren 1997; Savage et al. 2013). Nevertheless, in practice, SES is often viewed as a "shorthand expression" for variables indicating certain aspects of SES such as income, education, and occupation (Hauser and Warren 1997). These variables typically appear among standard demographic variables included in surveys, making it convenient to link SES to various other measures used in social science. SES is thus often measured or represented by one or a combination of these variables. The popular approaches to measure SES include using a univariate proxy such as just income or just education, a composite measure that incorporates income, education, and occupation such as Duncan's Socioeconomic Index (Duncan 1961) and the Nam-Powers occupational status scores (Nam and Powers 1965), or an occupation-based class schema such as the Erikson–Goldthorpe–Portocarero class schema (Erikson and Goldthorpe 1992).

Therefore, the most obvious approach to infer Twitter users' SES would be to estimate their income, education, or occupation. For instance, researchers can automatically extract job titles from users' profile description, rely on some sort of human validation to exclude inaccurately labeled jobs and then link the titles to income or occupational class (Ghazouani et al. 2019; Sloan et al. 2015). One can also obtain occupation from the LinkedIn links users include in their profile or tweets (Abitbol, Fleury, and Karsai 2019). The problem is that very few users state their job title or include a link to their professional accounts in their profile descriptions. Thus, the approach severely reduces the size of the sample to tens of thousands at most and potentially biases it toward individuals who act in official capacity, such as journalists, promoters, or politicians.

Using another data mining approach, researchers can estimate income or wealth by linking geo-located accounts and tweets to average house value or income at the census block level (Abitbol et al. 2019; Park et al. 2018). Similarly, however, users who disclose their geo-location are rare (Jiang et al. 2019). Around 30–40 percent of Tweets contain some profile location information, but the profile location tends to be at the region, state, city, or county level; the more granular geo-tagged tweets only make up one to two percent (Twitter 2022).

Employing more sophisticated machine learning techniques, other studies estimate SES with supervised methods trained on various Twitter features (Ghazouani et al. 2019). However, stemming from computer science, these studies do not engage sufficiently with social theory to justify the features and outcome variables used in the models (e.g., Filho et al. 2014; Moseley, Alm, and Rege 2014; Preoţiuc-Pietro, Lampos, and Aletras 2015; Volkova and Bachrach 2015; Volkova, Bachrach, and Durme 2016). For example, in one of the most cited papers on estimating Twitter users' SES, Preoţiuc-Pietro, Volkova, et al. (2015) employ the Bayesian nonparametric framework of Gaussian Processes to predict user income and occupational class from a large bag of features, including the number of followers, proportion of retweeted tweets, and the average number of tweets per day, among others, together with psycho-demographics, emotions, and word topics inferred from textual analysis of the user's posts. The authors train their model on the income and occupational class associated with the job titles retrieved from user descriptions. However, because they use too much information in estimating the SES with complex models, there is limited usage for the estimates. The paper also relies on aggregate-level information (income associated with job titles) to estimate individual SES without individual-level validation; this is another prevalent problem in the existing literature (e.g., Aletras and Chamberlain 2018; Ardehaly and Culotta 2017; Filho et al. 2014).

We contribute to existing efforts to estimate individual SES on Twitter by proposing an alternative unsupervised learning method. Political scientists have successfully used this method to estimate Twitter users' political ideology (Barberá 2015; Barberá et al. 2015) and here, we adapt it to estimate SES. The method relies on CA, a simple dimensionality-reduction technique that is already familiar to cultural and Bourdieusian sociologists, and is thus more accessible to less methodologically savvy social scientists than alternative complex supervised machine learning approaches such as Bayesian Gaussian Processes (Preoţiuc-Pietro, Volkova, et al. 2015) or neural graph embeddings (Aletras and Chamberlain 2018). The method uses minimal, commonly available, and easily accessible information about Twitter users' followings and employs fast off-the-shelf estimation algorithms, making it data economical, computationally efficient, and scalable. Specifically, the method yields SES estimates for millions of users compared to prior studies' benchmarks in the neighborhood of 50,000 (Aletras and Chamberlain 2018, Sloan et al. 2015). Finally, as we argue in the next section, the method relies on assumptions that are firmly embedded in classical sociological theory: Bourdieu's (1984) habitus theory. This renders the method relevant and useful for various strands of sociological research; it also directly responds to the recent call for better integration of data, measurement, and theory in CSS (Lazer et al. 2021; Wagner et al. 2021). Parenthetically, the proposed method aligns with the latest budding approaches to studying SES and culture with graph embeddings (Kozlowski, Taddy, and Evans 2019; Taylor and Stoltz 2020), as recent research shows the mathematical and interpretive similarity between CA and embedding methods (van Dam et al. 2021).

Measuring SES as Economic and Cultural Capital with Cultural Interests and Consumer Preferences

Bourdieu (1984) viewed an individual's SES as a function of their economic, cultural, and social capital. Economic capital refers to material resources such as wealth and income, cultural capital refers to the valued competence of engaging with cultural goods, and social capital refers to the network of contacts and connections that could be useful when needed. People's social position and the capital they possess shape how they act in and perceive the social world. Bourdieu calls this sense of orientation toward the social world *habitus*. The habitus manifests itself in people's everyday social practices and becomes concretely visible in people's cultural tastes and preferences. This manifestation may not be necessarily conscious and intentional but is nevertheless strategic, in the sense that it serves to distinguish one's social status and to distance oneself from other groups (Bourdieu 1984). Thus, on the one hand, people's upbring, education, and social surroundings shape their taste and cultural interests to be coherent within their own SES group. On the other hand, the exclusive nature of taste, which rejects cultural interests that are inconsistent with one's own SES, divides people into distinct and divergent SES groups.

Bourdieu mainly focused on the role of cultural tastes and cultural consumption in social distinction. Veblen ([1899] 2017) made a similar argument about distinction but instead emphasized the role of economic purchases. Using the concept of conspicuous consumption, Veblen argued that people tend to use material goods and leisure activities to demonstrate their SES to others. In other words, distinction could materialize not only via cultural tastes but also in preferences for consumer products and brands.

Naturally, Bourdieu's theory has been challenged, qualified, and extended since then. Most notably, while Bourdieu identified an accentuated taste stratification and classification in France, others have shown that, in the United States for example, individuals of higher social status tend to be "cultural omnivores," espousing broader and more eclectic cultural tastes (Holt 1998; Peterson 1992). Similarly, the recent notion of inconspicuous consumption suggests that people with more wealth and cultural capital actually tend to be more subtle and less ostentatious consumers (Berger and Ward 2010; Eckhardt, Belk, and Wilson 2015). Thus, more recent research challenges the idea that low versus high SES can be neatly mapped onto low- versus high-brow cultural tastes and basic versus luxury consumption. Nonetheless, it leaves intact two main assumptions that are crucial for our argument here: (1) people express their SES via their cultural interests and consumer preferences, and (2) people in similar SES tend to have similar cultural interests and consumption preferences.

Consequently, we argue that we can use the cultural interests and consumer preferences people declare on social media to estimate their SES. Specifically, we assume that Twitter users manifest their economic and cultural interests with the accounts they follow on Twitter. Many commercial and entertainment brands, including retailers (supermarkets, department stores, apparel), chain restaurants, news sources, sports associations, and TV shows, have official Twitter accounts. The brands use these accounts to share news, promote products and events, and interact and engage with fans, and users who value this information are more likely to follow these accounts. Marketing research shows that 35 percent of Twitter users in the US use Twitter to follow brands (Werliin 2020). Academic research shows that the main motivations for Twitter users to follow brands are incentives (discounts, coupons, promotions, etc.), information (to know more about products), social interactions (to interact with brand representatives or like-mined people), and attitudes toward brands (Kwon et al. 2014). These motivations align well with the framework of the *habitus*: preferences, interests, interactions, and attitudes represent different aspects of a person's orientation toward the social world, which reflects their socioeconomic background. Following consumer brands (e.g., retailers and chain restaurants) represents a combination of economic and cultural preferences: the price tag of the good or service reflects the economic constraints a person faces, and the associated quality and style represent the person's cultural taste and lifestyle. Following media and entertainment brands (e.g., news sources, sports associations, and TV shows) mainly represents cultural interests. Even if we do not know which brands represent higher economic and cultural capital, we can cluster users who tend to follow similar brands and project them onto a line, which will serve as our SES scale. This is the basic idea behind the method we propose below.

As a matter of fact, Bourdieu himself used a similar idea and a related method to demonstrate his concept of multidimensional social space. In his influential book *Distinction* (Bourdieu 1984), Bourdieu applied a dimensionality-reduction technique known as multiple CA (MCA) on a survey sample of the French population in the 1960s containing data on income, occupation, and engagement in various cultural activities (e.g., reading, going to concerts, and visiting museums). The technique allowed him to position individuals, occupations, and cultural activities on a two-dimensional graph. Bourdieu argued that the first dimension represents the overall volume of economic and cultural capital, and the second dimension represents the contrast between economic and cultural capital (Bourdieu 1984; Weininger 2005). Despite ongoing debates on the measurement of cultural capital and the relation between cultural interests and SES (Peterson and Kern 1996; Prieur and Savage 2013; Reeves 2019), a recent study reaffirmed the utility of using Bourdieu's method to establish social space and measure SES as a combination of economic and cultural capital (Flemmen, Jarness, and Rosenlund 2018).

In contrast to Bourdieu's surveys, we rely on the economic and cultural interests people reveal on social media. Computer scientists, political scientists, and psychologists have already used these data to extract various information about online users: demographic characteristics, political ideology, and psychological traits, as well as other private and sensitive information (Hinds and Joinson 2018). For instance, the researchers behind the myPersonality study apply supervised learning methods on participants' "likes" for Facebook groups to show that sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender can be predicted with relatively high levels of accuracy (Bachrach et al. 2014; Bi et al. 2013; Kosinski, Stillwell, and Graepel 2013; Youyou, Kosinski, and Stillwell 2015). More relevantly for us, political scientists utilize an unsupervised learning method to infer users' position on the left-right ideological spectrum based on the Twitter accounts of politicians, political parties, media outlets, and journalists the users follow (Barberá 2015; Barberá et al. 2015) or the Facebook pages of politicians they like (Bond and Messing 2015). The method uses CA (which is related to Bourdieu's MCA) on the users and the official accounts they follow to project their position on a continuous linear scale. Below, we outline how the method can be adapted to estimate user SES.

Method

The method relies on two sets of social media users: the accounts, public pages, or fan groups of consumer brands and cultural products and the individuals who follow, subscribe, or otherwise positively engage with them. It uses CA to map the associations between the brands and users onto a two-dimensional space and then estimate the SES of the brand/user from its coordinates in the first dimension. Based on our theoretical framing, we assume that the prime reason for a user to follow a brand is SES proximity, in the sense of congruent economic preferences, cultural interests, and lifestyle. Therefore, the first dimension from CA that explains the most variance of the user-brand matrix is a valid representation of the users and brands' SES. The use of CA is identical to political science approaches for estimating political ideology from Twitter followings and Facebook page likes (Barberá et al. 2015; Bond and Messing 2015) and in principle similar to the MCA conducted by Bourdieu himself (Bourdieu 1984; Flemmen et al. 2018).

CA is a multivariate method to summarize and visualize the associations between rows and columns of a two-way contingency table as the positions between points in a low-dimensional space (Greenacre 2017). The low-dimensional space is identified so that the variance of the original matrix is explained by the dimensions in descending order. Since the first two dimensions explain most of the variance, the output of CA is often a two-dimensional plot. In our case, we use the first dimension to obtain measures on a continuous SES scale but the method could be adapted to use the first two dimensions and assign SES according to a discrete class-based schema.

For **N** representing a binary matrix with *I* users as rows following *J* brands as columns, CA is conducted through the following main steps (Greenacre 2017).

First, we compute the matrix **S** of standardized residuals:

$$\mathbf{S} = \mathbf{D_r}(\mathbf{P} - \mathbf{rc})\mathbf{D_c}$$

where $\mathbf{P} = \frac{1}{\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}}\mathbf{N}$ is the binary data matrix transformed into proportions, **r** and **c** are the row and column weights with $r_i = \sum_{j=1}^{J} P_{ij}$ and $\mathbf{c_j} = \sum_{i=1}^{I} P_{ij}$, and $\mathbf{D_r} = \mathrm{diag}\left(1 / \sqrt{\mathbf{r}}\right)$ and $\mathbf{D_c} = \mathrm{diag}\left(1 / \sqrt{\mathbf{c}}\right)$ are the diagonal matrices with diagonal entries equal to the inverses of the square roots of the weights. This step ensures the model captures the associations between rows and columns in a way that does not depend on row or column sums. In essence, it accounts for the fact that some users are more active and some brands are more popular in general.

Second, we calculate the singular value decomposition of **S**:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_{\alpha}\mathbf{V}^{\mathrm{T}}$$

where **U** and $\mathbf{V}^{\mathrm{T}}$ are the left and right singular vectors of *S*, which are orthogonal and hence $\mathbf{U}\mathbf{U}^{T} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, and $\mathbf{D}_{\alpha}$ is the diagonal matrix of singular values in descending order $(\alpha_1 \geq \alpha_2 \geq \cdots)$. In other words, we now represent the information in **S** with two coordinate matrices (**U** and $\mathbf{V}^{\mathrm{T}}$) and a scaling matrix ($\mathbf{D}_{\alpha}$). Put simply, this step finds the low-dimensional space that best fits the original matrix in terms of least-squares approximation.

Finally, we project all rows and columns onto the plane by computing the standard coordinates: $\mathbf{G_r} = \mathbf{D_r} \mathbf{U}$ for rows and $\mathbf{G_c} = \mathbf{D_c V}$ for columns. As the original data matrix $\mathbf{N}$ lists users in rows and brands in columns, the row coordinates $\mathbf{G_r}$ in the first dimension give the estimated SES of the users, and the column coordinates $\mathbf{G_c}$ in the first dimension—the estimated SES of the brands. Lastly, we standardize the estimates to have a normal distribution with a mean of 0 and a standard deviation of 1, which aids the interpretation of the estimation. Since CA captures the relative positions of the users and brands, the interpretation of the estimated SES should focus on the values relative to other values in the whole sample rather than the absolute values. For example, an estimated user SES of -1 means that the user has an SES that is one standard deviation lower than the average user SES in the sample.

We note that CA also allows projecting data points (users or brands) not used in the original estimation onto the same subspace. To do this for a new brand, for example, we take the standardized column with the users that follow it $\mathbf{n}' = \frac{\mathbf{n}}{\sum_{i=1}^{I} \mathbf{n_i}}$ and compute $g = \mathbf{n}'^{\mathbf{T}}\mathbf{G_r}$. Similarly, we can map new users (Barberá et al. 2015).

Data

To test the validity of this method, we use the official Twitter accounts of a group of consumer brands and the followers of these accounts. Data collection and research for the study were approved by the LSE Ethical Review Board and the complete list of brands and their Twitter accounts required to replicate the results is available in Supplemental Table 1.

To identify the brands, we first selected six domains that cover various forms of daily material and cultural consumption: supermarkets and department stores, clothing and speciality retailers, chain restaurants, newspapers and news channels, sports, and TV shows. We then used Wikipedia lists, YouGov popularity rating lists, and media reports on TV shows' audiences (Maglio 2016, 2018; Wikipedia 2020; YouGov 2018) to identify the most prominent brands in the US. From these, we selected the ones that have a Twitter account with more than 10,000 followers. We only included accounts with a large number of followers to ensure the accounts can contribute to the analysis. Further, for international brands, we included only their US accounts, whenever available. We thus started with 341 brands.

Using the Twitter Search API (Twitter 2020) and the wrapper function in R developed by Barberá (2013/2020), we then obtained the full list of followers for these 341 brands till May 2020, yielding 191,790,786 users who follow at least one of the brands. To guarantee that we have sufficient information to characterize a user, we excluded users who follow fewer than five brands, which resulted in 23,567,268 users. Next, we used the users' profile data to further delete inactive users and potential bots. We kept users who had sent at least 100 tweets, have at least 25 followers, and had sent at least one tweet in the first five months of 2020. This selection left 4,436,095 users.

Finally, we were able to exclude some users who are not in the US based on the "location" field of their Twitter profile. We opted to exclude, rather than include users based on location data because these data are inconsistent and rarely available. For users who provide their location, some are easily identified just using text selection, as they put in a country or state name. For those who only put a street or city location, we used the Google Geolocation API (Google Developers 2020) to match the street or city with the country. After excluding users whose location is not in the US, there are 3,482,657 remaining users. After pruning the users, two brands ("Red Mango" and "Saatva Mattress") were left with only 0 and 1 followers, while the other brands had at least 1000. Since these two brands would not be informative for the analysis, we deleted them and then selected the users who follow at least five brands in the new sample. In the end, the sample contains a matrix of 339 brands and 3,482,652 users.

To improve the validity of the estimates, we conduct the analysis in two steps. First, we use CA on a maximally informative subset to identify the low-dimensional space and then, we project all users and brands to the space to estimate everyone's SES. Specifically, for the first step, we select "informative users" who follow at least one brand from each of the six domains (supermarkets and department stores, clothing and speciality retailers, chain restaurants, newspapers and news channels, sports, and TV shows), resulting in 158,441 users. Then we select the "informative brands" followed by at least 1000 of the "informative users," resulting in a 158,441 × 303 matrix (in comparison, the full matrix is 3,482,657 × 339).

We conduct CA on this subset using the *ca* package in R (Nenadic and Greenacre 2007). After confirming that the results are interpretable with a simple qualitative check, we use them to first project the coordinates for the rest of the brands, and then project the coordinates for the rest of the users. We use code from Barberá (Barberá [2013] 2020; Barberá et al. 2015) to do the projections.

Results and Validation

Figure 1 depicts the density distributions of the estimated SES for the brands in our sample and the users who follow them on Twitter. The estimated SES for the brands ranges from −2.95 (*hushpuppies_usa*) to 1.85 (*soulcycle*), with a median of 0.036. For the users, the estimated SES ranges from −7.00 to 2.02, with a median of 0.183. It is evident that both distributions are skewed toward middle-to-high SES. The skew for individuals corresponds well with the results from the nationally representative survey by Pew Research Centre showing that Twitter users are more educated and have higher income than the general US population (Wojcik and Hughes 2019).

To validate the estimates, we bring in data from various sources and conduct analyses at both the aggregate and individual levels. Our first step is to establish convergent validity. First, we confirm the qualitative interpretation of the brands' SES and compare our estimates with aggregate statistics on the educational level of the brands' marketing audience obtained from Facebook. Second, we quantify the extent to which, on aggregate, the SES estimates correlate with the mean salary and occupational class for a subsample of users who include an occupational title in their Twitter profile information. Third, we estimate the extent to which the individual SES estimates predict education and income in a small survey sample of Twitter users. Our next step is to confirm divergent validity, namely, that the SES estimates are not measuring other related demographic variables. We use again data from Facebook and Twitter users, and the survey sample to confirm that the SES estimates are to a much lesser extent associated with age, gender, race, political ideology, and urban/rural residence.

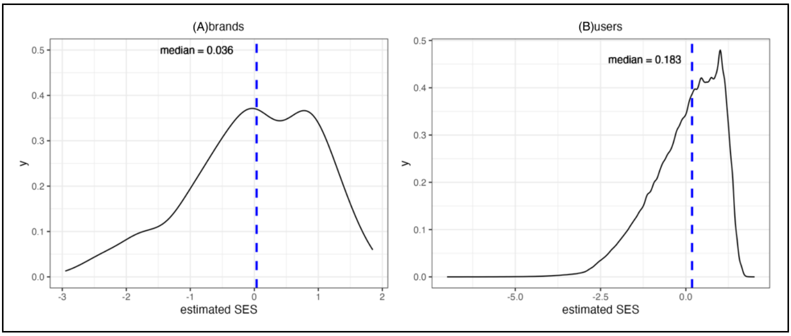

Figure 1. Density plot of the estimated socioeconomic status (SES) for (A) 339 brands and (B) 3,482,652 users who follow them on Twitter.

We note that since SES is a composite concept, and our measure is operationalized to capture this multifacetedness, we do not expect a perfect correlation between our SES estimates and any single one of the simple measures of education, occupational class, or income. Yet, neither can we rely on another composite measure such as the SEI as a ground-truth benchmark to measure our success against: as we mentioned in the introduction, the sociological community has not coalesced to a universal understanding of SES. Our primary aim here is to prove the existence of a meaningful signal in the proposed measure and stimulate further research that could better isolate, filter, and amplify this signal.

Validation of Brand SES

We begin by qualitatively sense-checking the SES estimates for brands. Figure 2 shows the estimates for a selected group of popular brands, while Supplemental Table 2 lists all estimates. The lower end of the scale has discount store chains such as *Family Dollar, Dollar General,* and *True Value*. Slightly higher, there are fast food restaurant chains such as *Pizza Hut, KFC,* and *Burger King*, and inexpensive stores and supermarket chains such as *Big Lots* and *Aldi*. The next band, constituting the first hump of the bimodal distribution visible in Figure 1A, contains many essential and/or large businesses: *McDonald's*, *Walmart*, *Best Buy*, *Home Depot*, *Old Navy*, and *Toys "R" Us*. Then, there are average priced supermarket and clothing brands such as *Target*, *H&M,* and *Gap*. The most populated SES band (the second peak in Figure 1A) has the brands that one could argue are universally popular, such as *Nike* for clothing, *NFL* and *NBA* for sports, *the Big Bang Theory* for TV shows, and *Starbuck*s for coffee chains. The higher end has iconic middle to elite-class brands such as *Whole Foods,* chic and expensive exercise brands *Peloton* and *Soul Cycle,* and the TV show *Mad Men*, which in

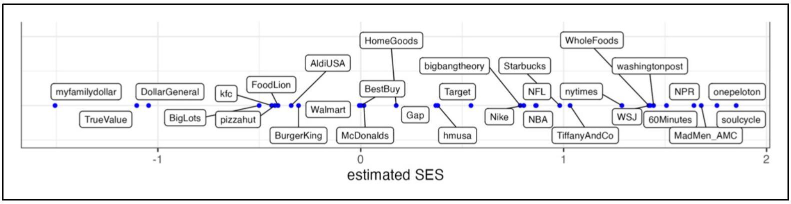

Figure 2. Estimated socioeconomic status (SES) for a selected group of popular brands.

2010 was reported to have 48 percent of its viewers with a household income of more than $100,000 (Szalai 2010). The higher end also includes national newspapers such as *The New York Times*, *The Wall Street Journal,* and *The Washington Post*. This result corresponds well with Chan and Goldthorpe's research (2007b), which shows that national newspapers tend to be read by people with higher social status.

In the next step, we validate the brands' estimated SES quantitatively with data from the Facebook Marketing API (Facebook 2021). Prior research on migration, health, urban crime, and digital inequalities (e.g., Araujo et al. 2017; Fatehkia et al. 2018) demonstrates that the Facebook Marketing API can be an effective tool for obtaining population-level demographic estimates. With tailored targeting criteria, the API provides the number of users an ad can reach per month on Facebook. We use the targeting criteria to choose an interest, for example, *soulcycle*, and find the number of active users whose highest earned degree is high school diploma, Bachelor's degree, and Master's or higher and who express interest in *soulcycle* in the US, from which we then calculate the proportion of *soulcycle*'s audience with different educational levels. We recognize that the audience on Facebook and Twitter is not entirely the same; expressing interest in a brand on Facebook and following a brand on Twitter may also represent different motives. Nonetheless, the Facebook audience data provide valuable insights into the brands' audience composition and thus offer a useful reference for the validation of our measurement.

There are multiple educational levels in the Facebook data, including categories such as "in university" and "some degree." For clarity, we only choose three levels that represent the full completion of a degree. Seven brands (*FinishLine, GNCLiveWell, GreysABC, Gap, LEVIS, MakitaTools*, and *CodeBlackCBS*) in our sample have an audience size of 1000 universally in all educational levels, which may mean Facebook does not have a reasonable estimate of the audience size for these brands. Further, no suitable data are available for four brands (*moen, Hanes, thehill,* and *WestworldHBO*). Therefore, we exclude these brands for this part of the analysis, resulting in 328 brands. Figure 3 plots the proportion of the brand's Facebook audience at the specified educational level against the brand's estimated SES according to our method. A small number of the brands' Twitter screen names are shown alongside their points and to aid visibility, these are chosen for plot areas with low density of observations. Panel A shows a negative association between the brand's estimated SES and the proportion of users in the brand's Facebook audience whose highest earned degree is a high school diploma (Spearman's $\rho = -0.464$, $p < .001$), while panel C shows a positive

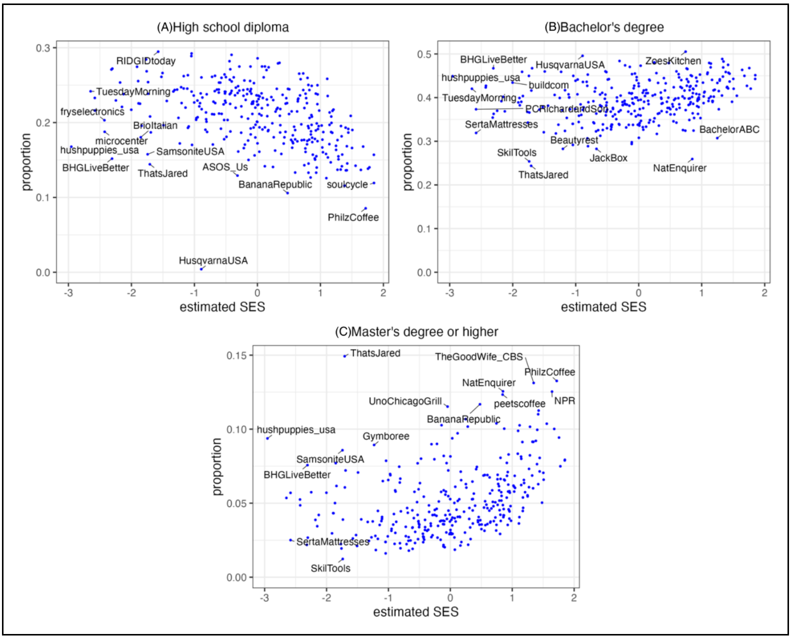

Figure 3. Relation between the educational composition of the brands' Facebook audience, as measured by the proportion who have earned at most the respective accreditation, and the brands' estimated socioeconomic status (SES).

association between the estimated SES and the proportion who hold a Master's or higher degree ($\rho = 0.444$, $p < .001$). Panel B shows a somewhat lower but still positive association between the estimated brand SES and the proportion of users among the brand's audience whose highest degree is Bachelor's ($\rho = 0.320$, $p < .001$).

The patterns in the plots and the correlation statistics show that the brands with higher estimated SES tend to have a significantly smaller audience of at-most high school graduates, a significantly larger audience with Master's or higher degrees, and a somewhat larger audience of Bachelor degree holders. The latter represents the largest and most diverse audience on Facebook, so it is expected that their expressed interests in the brands are not as informative as the other education levels. These trends together suggest that the audience of the brands with higher estimated SES has a higher average educational level than the audience of the brands with lower estimated SES. In sum, the proposed method positions consumer and media brands along an SES scale in ways that resonate with common knowledge and convincingly capture the educational level of the brand's social media audience.

Validation of User SES with Self-Reported Job Titles

We next assess whether the SES estimates for users are valid too, starting at the aggregate level. We do this by identifying a set of common and informative job titles mentioned in Twitter profiles and comparing the income and occupational class associated with the job title to the average SES estimates for the Twitter users who state this job title in their profile description. Essentially, we quantify how the estimates by our SES measurement method compare on average to those by another prominent approach that relies on self-disclosed job titles (Sloan et al. 2015).

We complete the following steps to identify and match job titles. We first find job titles from different occupational social classes from the UK's Standard Occupational Classification (ONS 2020) and note their class. We choose the UK's SOC instead of the US's SOC because it has more specific job titles and is closer to the well-established Goldthorpe Class Scheme (Goldthorpe, Llewellyn, and Payne 1987; Rose, Pevalin, and O'Reilly 2005). Then we use text selection to search for the job titles in the profile descriptions of all users in our Twitter sample. We only include the job titles that return more than 50 users. To minimize the number of wrongly labeled titles, we include an additional filter: we manually inspect ten randomly sampled descriptions for each job title to identify text structures that contribute to mislabeling and then filter out the titles that match the text structures identified. After this filtering, we also delete two titles (tailor and waitress) that have fewer than ten cases. In the 2020's version of the UK SOC scheme, there are nine occupational social class levels, where a lower number means a higher occupational social class (ONS 2020). We try to include job titles from all nine classes, but job titles in some classes are harder to match with profile descriptions than others. After the text selection, we search the job titles in the "May 2019 National Occupational Employment and Wage Estimates" table on the website of the US Bureau of Labour Statistics (2020). We only include job titles that make sense in the US context and note their mean annual salaries. The outlined procedure resulted in a sample of 42,099 users matched with 50 titles, which we use as our validation set. Supplemental Table 3 lists the selected titles and their mean annual salary in US dollars, grouped by their occupational social class.

Figure 4 depicts the association between the median estimated SES of users for each job title and the job title's mean annual salary and occupational class. The salary is logarithm scaled with base 10, the bars show standard errors for the median estimated SES, and the colors and shapes represent the occupational social class, where higher number means lower class. There is a clear positive trend, where jobs with a higher median estimated SES tend to have a higher mean annual wage. Jobs with the same class also tend to cluster. From bottom left to top right, there is a discernible trend from low salary, class, and estimated SES to higher salary, class, and estimated SES. Statistical tests show that Spearman's rank correlation between the median estimated SES and mean annual salary is 0.673 ($p <$ .001). The Spearman's rank correlation between the median estimated SES and occupational class is −0.640 ($p <$ .001). As a reference, in our sample, Spearman's correlation between mean annual salary and class is −0.840 ($p <$ .001). Although the correlations between our estimated SES and salary or class are not as high as the well-established correlation between salary and class, they are sufficiently strong to validate the proposed method at the aggregate level.

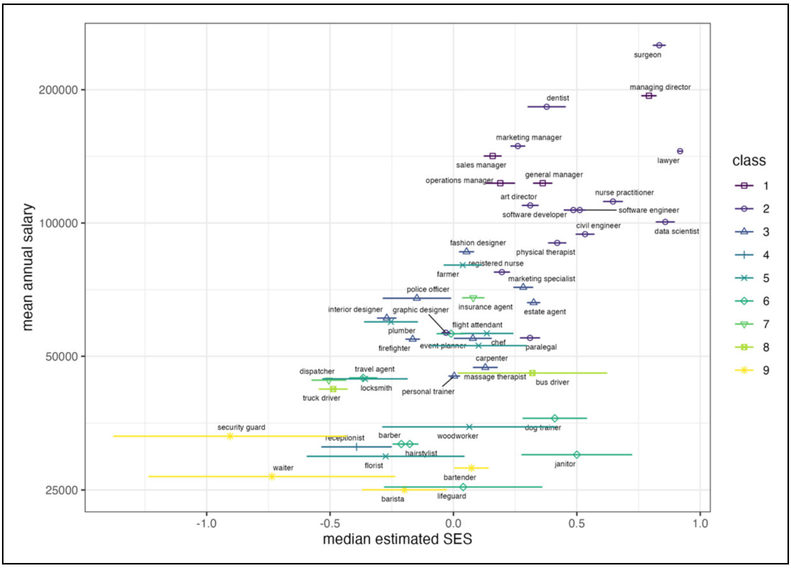

Figure 4. Relation between median estimated socioeconomic status (SES), mean salary, and occupational class for a set of 50 common job titles, estimated over 42,099 Twitter users who mention one of the titles in their profile description.

Nevertheless, as the error bars in Figure 4 show, there are large variations of estimated SES for some job titles, especially at the lower end of SES. Individual earnings for the same job title vary depending on US state, urban setting, business size, etc. At the aggregate level, the effects of these variations may cancel out by the large number of users selected for each title, but the effects will be more palpable at the individual level. Therefore, we next use individual-level SES data to further validate our estimates.

Validation of User SES with Survey Data

To test the method with better ground truth data, we identify a small sample of Twitter brand followers who report their income and educational level in a survey. The survey data were provided by Guess et al. (2021), who recruited 1,551 respondents from the YouGov U.S. Pulse panel, 471 of whom shared their Twitter data. Restricting the sample to users who follow at least one of the brands from our sample, we were left with 200 users whose SES we can estimate with our method. For these 200 users, we have their self-reported highest educational level as ordinal variable from one to six, coded as 1: No high school, 2: High school graduate, 3: Some college, 4: Two-year college, 5 Four-year college, and 6: Post-graduate. For 182 users, we also have their income coded as an ordinal variable from one to 16, starting from less than \$10,000, then going in increments of \$10,000 up to \$80,000, after which the categories start from \$100,000, \$120,000, \$150,000, \$200,000, \$250,000, \$350,000, and finally, \$500,000 or more. Using these data, visually presented in Figure 5, we estimate the Spearman correlation between estimated SES and educational level to be 0.269 ($p < .001$) and the one between estimated SES and income level to be 0.188 ($p < .05$). As a reference, the Spearman correlation between income and education in the sample is 0.455 ($p < .001$), which is surprisingly low. If we restrict the survey sample to Twitter users who follow at least two or three accounts, the correlations with education improve (0.259, $p < .001$, $N = 147$ in the case of two accounts; 0.344, $p < .001$, $N = 111$ for three accounts) but weaken for income (0.137, $p = .117$, $N = 131$ for two accounts; 0.156, $p = .117$, $N = 102$ for three accounts). These results suggest that our method successfully captures information relating to SES and specifically, captures education better than income. Figure 5 reveals that the model is particularly successful in identifying highly educated individuals with high income. Nevertheless, there appears to be a significant amount of noise or, possibly, unrelated demographic information. Ideally, we would have access to larger survey data to

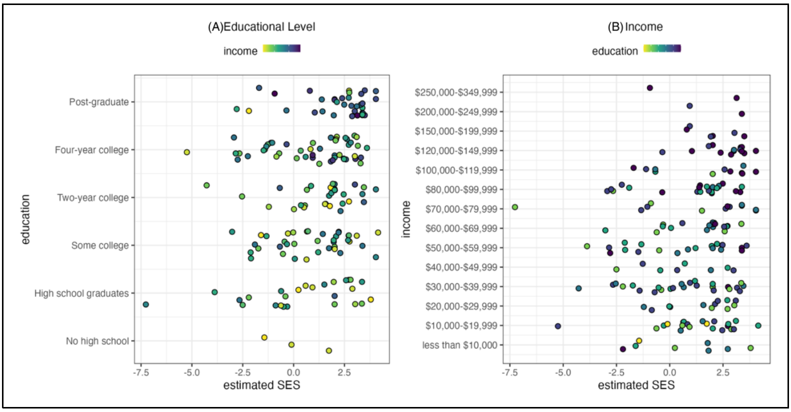

Figure 5. Relation between estimated socioeconomic status (SES) and (A) educational level and (B) income for 200 (182 for B) survey respondents who follow at least one of the 339 brands on Twitter. The y-axis values are plotted with noise to improve visibility.

identify for whom the method underperforms. At the very least, we should establish that the proposed method captures SES constructs better than other associated demographic variables. This is what we do next.

Divergent Validity

So far, we focused on convergent validity, utilizing multiple sources of data to establish that the estimates are correlated with other proxies for the theoretical concept of SES. To further establish the validity of the measurement method, we also provide evidence for divergent validity, demonstrating that the estimated SES does not capture other demographic variables related to SES better.

First, with similar data from the Facebook Marketing API, we analyze the associations between the estimated SES and the proportion of urban users, male/female users, and users in different age groups.[1] The estimated SES of the brands is very weakly associated with the proportion of urban users ($\rho = 0.114$, $p = .050$) and not associated with the proportion of male ($\rho = 0.034$, $p = .558$) nor female ($\rho = -0.037$, $p = .532$) users. These results suggest our SES measure for the brands is not capturing urban/rural or gender disparity. The estimated SES of the brands has significant but weak positive associations with the proportion of users in younger age groups:

Spearman correlation coefficients of 0.172 ($p < .01$) for age 18–24, 0.199 ($p < .001$) for 25–34, and 0.136 ($p < .05$) for 35–44. Conversely, the estimates have weak negative associations with older age groups: Spearman correlation coefficients of −0.135 ($p < .05$) for age 45–54, −0.224 ($p < .001$) for 55–64, and −0.117 ($p < .05$) for 65 and above. Although statistically significant, the associations between estimated SES and age are much weaker than education and hence, we can conclude that the estimated SES for the brands captures education better than age.

Second, we test the correlation between estimated SES and political ideology, as measured by Barberá et al.'s (2015) method. For the 150,011 informative users whose Twitter followings are still available in November 2022, the Spearman correlation between estimated SES and political ideology is −0.114 ($p < .001$), where positive values for ideology mean conservative-leaning. The large sample size contributes to the statistical significance, but the correlation is weak. Thus, although we use the same method and several overlapping official Twitter accounts (mainly news), the modification we propose no longer reflects political ideology at the individual level.

Third, using again data from Guess et al. (2021), we test the associations between estimated SES and related demographic variables available in the survey: age, gender, political ideology, and race. The estimated SES is not significantly associated with any of the variables tested. For the 195 participants with available demographic data, the Spearman correlation with age is 0.106 ($p = .139$) and the t-test between male and female is 0.741 ($p = .460$). Similarly, there is no significant difference in estimated SES between the four racial groups (White, Black, Hispanic, and Asian/other) categorized in the survey, regardless of whether we use an analysis of variance test ($p = .871$) or pair-wise t-tests. For the 189 participants with self-reported political ideology (a scale from 1 to 5), the correlation between estimated SES and political ideology is −0.079 ($p = .279$). As a reference, in the sample, education and income are also not significantly associated with any of the four variables (detailed results are available in Supplemental Table 5). Further, regression analyses, presented in Supplemental Table 6, show that controlling for age, gender, political ideology, and race, there are still significant correlations between estimated SES and education ($p < .001$) and between estimated SES and income ($p < .05$).

Overall, the estimated SES has insignificant or weak associations with related demographic variables such as age, gender, race, political ideology, and urban/rural residence, while the correlations between the estimated SES and established SES proxies, including education, income, and occupational class, are significant and much stronger. Combined together, the results of the analyses of convergent and divergent validity provide a strong case for the validity of the proposed method.

Discussion

This study presents a method for estimating Twitter users' SES from the consumer and media brands they follow. The method is adapted from a widely used approach to measuring Twitter users' political ideology. Compared to previous attempts to estimate SES from social media data, the proposed method is built on behavioral assumptions that can be linked to classical sociological theory, requires only a basic understanding of a common dimensionality reduction technique, and provides estimates for millions of individuals while only using minimal, easily available and obtainable data, open-source off-the-shelf software programs, and modest computational power. We applied the method using 339 popular US brands to estimate the SES of almost 3.5 million Twitter users. We then brought in additional data, including advertisement audience statistics from Facebook, user profile information from Twitter, and survey sample responses, to validate the accuracy of the estimates with the standard SES proxies of education, occupational class, and income and confirm their dissociation from other demographic variables known to be related to SES.

The results suggest that the proposed measure of SES for Twitter users is promising. The measure works well at the aggregate level but needs fine-tuning with better validation data for more precise individual estimates. The estimated SES for the brands correlates reasonably well with the educational level of their audience and aligns intuitively with general brand perceptions. Aggregated for a selected group of job titles, the estimated SES for users is also strongly associated with annual mean salary and occupational class. At the individual level, the SES estimates are significantly associated with education and income, but the correlations are relatively weak. Further, for both brands and individuals, the SES estimates are not, or at best much weakly, associated with related demographic variables, including age, gender, race, urban/rural residence, and political ideology. Overall, the significant associations between the estimated SES and the traditional SES indicators and the insignificant or weak associations with other demographic variables at both the aggregate and individual levels support the underlying principle of the proposed method and justify further efforts to refine it at the individual level.

Nevertheless, we interpret the results with some further reflections on the theoretical assumptions and methodological choices we made. The main principle of the proposed method is that Twitter users manifest their economic and cultural interests with the brands they follow and hence these brands can inform us about their SES. We note that following a brand on Twitter does not involve any economic costs and does not necessarily imply real material consumption. Yet, no economic cost does not mean no cost at all. Users have finite ability to process information and divide attention on Twitter (Hodas and Lerman 2012). Following an account populates one's newsfeed with updates, displacing other relevant information and this is particularly the case for official accounts managed by professionals who regularly produce content. In other words, while clicking to follow Whole Food's Twitter account is just as effortless as clicking to follow Aldi's account, there are direct and opportunity information costs associated with remaining a follower.

Unconstrained by cost, Twitter users may follow brands for many possible reasons that are not relevant to economic or cultural interests, for example, out of curiosity or by mistake. We certainly cannot assume that all brand followings are based on economic and cultural interests associated with SES, but we propose that the dominant trend is related to SES. The validation results indeed indicate that SES has a significant role to play. This observation also aligns with evidence that the digital world reflects and even reproduces the existing cultural boundaries of the physical world regarding people's interests in restaurants, music, films, museums, and galleries (Airoldi 2021; Goldberg, Hannan, and Kovács 2016; Mihelj, Leguina, and Downey 2019), and even more so, politics (Bail et al. 2018; Tucker et al. 2018). The basic principle behind the proposed method is to exploit these digital cultural and lifestyle boundaries to obtain information about individuals, which can then be used in research that challenges them.

Another related objection is that following a brand on Twitter might be aspirational and reflect desired, rather than actual SES. We know that, on the one hand, people universally desire higher social status (Anderson, Hildreth, and Howland 2015; Fiske 2011) and on the other, online users strategically orchestrate online personas and actively manage their self-presentation online (Schlenker and Pontari 2000). However, since followed accounts are not easily and directly observable on a user's profile, they are unlikely to be employed solely as status signals. A user can signal status with the accounts they follow only if they actively retweet or @-mention them, so future work could analyze such activity to estimate the extent to which followings are status-seeking rather than status-reflecting. Additionally, we note that the unsupervised learning method we employ is agnostic to *a priori* brand associations or expectations. The method positions the brands according to their co-followings and it can thus place an expensive brand toward the low-SES end of the spectrum if its audience on Twitter tends to consist of consumer-hopefuls rather than actual consumers. Nevertheless, we recognize that strategic self-presentation may be idiosyncratic and as such, it will inevitably introduce noise to the individual estimates.

Finally, we note that the weak signal at the individual level the method detects should be interpreted in light of the natural limits of predictability of human behavior social scientists face (Hofman, Sharma, and Watts 2017; Song et al. 2010). As we discussed above, besides actual SES, strategic self-presentation, unknown personal motivations, other demographic characteristics, peer effects, and situational factors could dictate whether a specific individual follows a brand. This inevitable degree of idiosyncrasy and complexity means that the salient effect of SES may only manifest at the aggregate level but dissolve at the individual level. A recent large-scale mass collaboration scientific project shows that, even with high-quality data and sophisticated methods, the predictability of individual life outcomes is still very low (Salganik et al. 2020). We soberly recognize that similar natural limits likely constrain the measurement of individual SES of Twitter users from their expressed cultural interests and consumer preferences.

Despite these inherent limitations, we see a huge potential in further efforts to validate, refine, and apply the proposed method. The next natural step is to link richer survey data of a larger sample with Twitter user profiles. This step involves extra resources and additional methodological and ethical issues (Baghal et al. 2021; Stier et al. 2020) but the resulting linked data could contribute in multiple ways. First, the data will allow for revalidating the proposed method, disentangling demographic factors that strongly influence the SES estimates, and quantifying the extent to which the measures correspond to actual versus desired SES. Second, the data can be used to fit supervised learning models for estimating SES to not only improve the proposed unsupervised method but also compare the strengths and weakness of different methods, examine the inherent limits to the predictability of individual SES, and recommend suitable methods for different situations.

One way in which a supervised learning model on linked survey data could help improve the proposed method is by refining the consumer domains and official accounts to include in the estimation. The included official accounts determine whether CA indeed captures the variations in SES. In this study, we consulted a variety of sources to select a group of brands that represent a wide range of economic and cultural interests, but this selection could be improved with a more data-driven approach. Although there are numerous studies on the link between taste and social status, especially following Bourdieu's (1984) work (e.g., Alderson et al. 2007; Chan and Goldthorpe

2007b; Gerhards et al. 2013; Katz-Gerro 1999; Peterson 1992; Reeves 2019), there is limited research on the specific brand preferences of people in different SES. The brands themselves rarely disclose their audience demographics. Future research would benefit from a comprehensive analysis of the relation between SES and specific interests using sources such as the Facebook Marketing API and other mobile or web tracking data, linking it to previous research on SES and taste. Such research will provide not only a more informed selection of the official accounts to include in the model but also a more comprehensive picture of SES, taste, and habitus.

Although we carefully considered the six domains we chose (supermarkets and department stores, clothing and speciality retailers, chain restaurants, newspapers, and news channels, sports, and TV shows), this set is not necessarily comprehensive. One may argue that news sources, sports, and TV shows are very reductive parts of cultural interests that people express on Twitter, and that artists, musicians, and influencers should also be included. Indeed, the current set of domains carries the danger of reducing cultural capital to consumerism, especially with its focus on "brands." For this initial attempt, we took a more conservative approach and chose consumer brands that combine economic and cultural interests, avoiding accounts related to art and music. Music and art not only form the core of cultural capital but also fuel intense debates about the link between cultural capital and SES. The highbrow-vs-omnivore debate around cultural capital, where art and music activities are often used as empirical evidence, is ongoing and active (Chan 2019a; Goldberg 2011; Peterson 1992; de Vries and Reeves 2022). We thus expect the contribution of musicians and artists to SES estimation in our method would be less informative and less interpretable. Nevertheless, this constitutes an empirically testable hypothesis that future work could explore. Work that adds artists, musicians, and other related accounts to the proposed model could potentially both benefit our method and contribute to the ongoing highbrow-vs-omnivore debate.

Despite the mentioned limitations and aspects for improvement, the proposed method carries an enormous promise for social science research. The method provides SES estimates on a continuous scale that are operationally easy to use and theoretically interpretable. Social scientists could combine these SES estimates with digital trace data on behaviors, communication patterns, and social interactions to study inequality, health, and political engagement, among many other topics. For instance, one can link our measure of SES, which captures cultural and economic capital, to indicators of social capital inferred from social relations and interactions on Twitter and explore how the different forms of capital combine to contribute to socioeconomic inequality. Specifically, we can now study the effects of social networks on inequality, as discussed by DiMaggio and Garip (2012), with new data, in a different context, and on a significantly larger scale.

The SES estimation method we propose here opens myriad new avenues for academic research on Twitter and similar social network platforms. We used Twitter due to its popularity and convenient API, but the principle of our method can be applied to many other online platforms. For example, future research can use the interests expressed by following or liking certain topics or accounts to estimate the SES of users on platforms such as Reddit and Quora and then link SES to behaviors, opinions, and knowledge expressed on those platforms.

Authors' Note

Data, code, and a README file detailing what is included and how to reproduce the published results are available on *Figshare* (https://doi.org/10.6084/m9.figshare.22007000.v1) and *GitHub* (https://github.com/yuanmohe/Twitter_SES).

Acknowledgements

The authors are grateful to Andrew Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang for sharing data, to Eleanor Power and Blake Miller for valuable detailed feedback, and to the reviewers for their constructive comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: M.T. received funding from the Volkswagen Foundation (Grant Ref. 92 173). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ORCID iDs

Yuanmo He https://orcid.org/0000-0001-7827-5395
Milena Tsvetkova https://orcid.org/0000-0002-3552-108X

Supplemental Material

Supplemental material for this article is available online.

Note

[1] The divergent validity analysis was conducted two years after the convergent validity analysis, during which period the Facebook Marketing API changed the searchable terms and some brands went bankrupt. Therefore, the number of brands with suitable audience data dropped from 328 to 295. The details of the unavailable brands are included in Supplemental Table 4. Additionally, the API now does not return one number for the estimated target audience size but returns lower bound and upper bound. Here, we present the results using the average between lower and upper bound. The results from the average, lower, and upper bound are essentially the same.

References

Abitbol, Jacob, Eric Fleury, and Márton Karsai. 2019. “Optimal Proxy Selection for Socioeconomic Status Inference on Twitter.” *Complexity* 2019:e6059673.

Abitbol, Jacob L. and Márton Karsai. 2020. “Interpretable Socioeconomic Status Inference from Aerial Imagery Through Urban Patterns.” *Nature Machine Intelligence* 2(11): 684-92.

Adler, Nancy E., Thomas Boyce, Margaret A. Chesney, Sheldon Cohen, Susan Folkman, Robert L. Kahn, and S. Leonard Syme. 1994. “Socioeconomic Status and Health: the Challenge of the Gradient.” *American Psychologist* 49(1):15-24.

Airoldi, Massimo. 2021. “The Techno-Social Reproduction of Taste Boundaries on Digital Platforms: the Case of Music on YouTube.” *Poetics* 89:101563.

Alderson, Arthur S., Azamat Junisbai, and Isaac Heacock. 2007. “Social Status and Cultural Consumption in the United States.” *Poetics* 35(2):191-212.

Aletras, Nikolaos and Benjamin P. Chamberlain. 2018. “Predicting Twitter User Socioeconomic Attributes with Network and Language Information.” Pp. 20–24 in *Proceedings of the 29th on Hypertext and Social Media, HT ‘18*. Baltimore, MD, USA: Association for Computing Machinery.

Anderson, Cameron, John Angus D. Hildreth, and Laura Howland. 2015. “Is the Desire for Status a Fundamental Human Motive? A Review of the Empirical Literature.” *Psychological Bulletin* 141(3):574-601.

Araujo, Matheus, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. 2017. “Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations.” ArXiv:1705.04045 [Cs].

Ardehaly, Ehsan M. and Aron Culotta. 2017. “Mining the Demographics of Political Sentiment from Twitter Using Learning from Label Proportions.” Pp. 733-38 in 2017 IEEE International Conference on Data Mining (ICDM).

Bachrach, Yoram, Thore Graepel, Pushmeet Kohli, Michal Kosinski, and David Stillwell. 2014. “Your Digital Image: Factors Behind Demographic and

Psychometric Predictions from Social Network Profiles." Pp. 1649-50 in Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems.

Baghal, Tarek Al, Alexander Wenz, Luke Sloan, and Curtis Jessop. 2021. "Linking Twitter and Survey Data: Asymmetry in Quantity and Its Impact." *EPJ Data Science* 10(1):10-32.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115(37):9216-21.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76-91.

Barberá, Pablo. [2013] 2020. "Pablobarbera/Twitter_ideology."

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting From Left to Right: is Online Political Communication More Than an Echo Chamber? ." *Psychological Science* 26(10):1531-42.

Berger, Jonah and Morgan Ward. 2010. "Subtle Signals of Inconspicuous Consumption." *Journal of Consumer Research* 37(4):555-69.

Bi, Bin, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. "Inferring the Demographics of Search Users: Social Data Meets Search Queries." Pp. 131-40 in Proceedings of the 22nd International Conference on World Wide Web, WWW '13. Rio de Janeiro, Brazil: Association for Computing Machinery.

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(1):62-78.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.

Brady, Henry E., Sidney Verba, and Kay L. Schlozman. 1995. "Beyond SES: a Resource Model of Political Participation." *American Political Science Review* 89(2):271-94.

Brandwatch. 2020. "60 Incredible and Interesting Twitter Stats and Statistics." *Brandwatch*. Retrieved 16 December 2020. https://www.brandwatch.com/blog/twitter-stats-and-statistics/.

Campbell, Karen E., Peter V. Marsden, and Jeanne S. Hurlbert. 1986. "Social Resources and Socioeconomic Status." *Social Networks* 8(1):97-117.

Chan, Tak W. 2019a. "Understanding Social Status: a Reply to Flemmen, Jarness and Rosenlund." *The British Journal of Sociology* 70(3):867-81.

Chan, Tak W. 2019b. “Understanding Cultural Omnivores: social and Political Attitudes.” *The British Journal of Sociology* 70(3):784-806.

Chan, Tak W. and John H. Goldthorpe. 2007a. “Class and Status: the Conceptual Distinction and Its Empirical Relevance.” *American Sociological Review* 72(4):512-32.

Chan, Tak W. and John H. Goldthorpe. 2007b. “Social Status and Newspaper Readership.” *American Journal of Sociology* 112(4):1095-134.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, Drew Johnston, Martin Koenen, Eduardo Laguna-Muggenburg, Florian Mudekereza, Tom Rutter, Nicolaj Thor, Wilbur Townsend, Ruby Zhang, Mike Bailey, Pablo Barberá, Monica Bhole, and Nils Wernerfelt. 2022. “Social Capital I: Measurement and Associations with Economic Mobility.” *Nature* 608(7921):108-21.

Choi, Hyunyoung and Hal Varian. 2012. “Predicting the Present with Google Trends.” *Economic Record* 88(s1):2-9.

de Vries, Robert and Aaron Reeves. 2022. “What Does It Mean to Be a Cultural Omnivore? Conflicting Visions of Omnivorousness in Empirical Research.” *Sociological Research Online* 27(2):292-312.

Diemer, Matthew A., Rashmita S. Mistry, Martha E. Wadsworth, Irene López, and Faye Reimers. 2013. “Best Practices in Conceptualizing and Measuring Social Class in Psychological Research.” *Analyses of Social Issues and Public Policy* 13(1):77-113.

DiMaggio, Paul and Filiz Garip. 2012. “Network Effects and Social Inequality.” *Annual Review of Sociology* 38(1):93-118.

DiPrete, Thomas A. and Gregory M. Eirich. 2006. “Cumulative Advantage as a Mechanism for Inequality: a Review of Theoretical and Empirical Developments.” *Annual Review of Sociology* 32(1):271-97.

Dohrenwend, B. P., I. Levav, P. E. Shrout, S. Schwartz, G. Naveh, B. G. Link, A. E. Skodol, and A. Stueve. 1992. “Socioeconomic Status and Psychiatric Disorders: the Causation-Selection Issue.” *Science (New York, N.Y.)* 255(5047):946-52.

Duncan, Otis D. 1961. “A Socioeconomic Index for All Occupations.” Pp. 109-38 in *Occupations and Social Status*. New York: Free Press.

Eagle, N., M. Macy, and R. Claxton. 2010. “Network Diversity and Economic Development.” *Science* 328(5981):1029-31.

Eckhardt, Giana M., Russell W. Belk, and Jonathan A. J. Wilson. 2015. “The Rise of Inconspicuous Consumption.” *Journal of Marketing Management* 31(7–8):807-26.

Erikson, Robert and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford, UK: Oxford University Press.

Facebook. 2021. “Marketing API - Documentation.” *Facebook for Developers*. Retrieved 22 January 2021. https://developers.facebook.com/docs/marketing-apis/.

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. 2018. “Using Facebook Ad Data to Track the Global Digital Gender Gap.” *World Development* 107:189-209.

Filho, Renato M., Guilherme R. Borges, Jussara M. Almeida, and Gisele L. Pappa. 2014. “Inferring User Social Class in Online Social Networks.” Pp. 1-5 in *Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD’14*. New York, NY, USA: Association for Computing Machinery.

Fiske, Susan T. 2011. *Envy Up, Scorn Down: How Status Divides Us*. New York, NY: Russell Sage Foundation.

Flemmen, Magne, Vegard Jarness, and Lennart Rosenlund. 2018. “Social Space and Cultural Class Divisions: the Forms of Capital and Contemporary Lifestyle Differentiation.” *The British Journal of Sociology* 69(1):124-53.

Flemmen, Magne P., Vegard Jarness, and Lennart Rosenlund. 2019. “Class and Status: on the Misconstrual of the Conceptual Distinction and a Neo-Bourdieusian Alternative.” *The British Journal of Sociology* 70(3):816-66.

Gerhards, Jürgen, Silke Hans, and Michael Mutz. 2013. “Social Class and Cultural Consumption: the Impact of Modernisation in a Comparative European Perspective.” *Comparative Sociology* 12(2):160-83.

Ghazouani, Dhouha, Luigi Lancieri, Habib Ounelli, and Chaker Jebari. 2019. “Assessing Socioeconomic Status of Twitter Users: A Survey.” Pp. 388-98 in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). Varna, Bulgaria: INCOMA Ltd.

Goldberg, Amir. 2011. “Mapping Shared Understandings Using Relational Class Analysis: the Case of the Cultural Omnivore Reexamined.” *American Journal of Sociology* 116(5):1397-436.

Goldberg, Amir, Michael T. Hannan, and Balázs Kovács. 2016. “What Does It Mean to Span Cultural Boundaries? Variety and Atypicality in Cultural Consumption.” *American Sociological Review* 81(2):215-41.

Goldthorpe, John H., Catriona Llewellyn, and Clive Payne. 1987. *Social Mobility and Class Structure in Modern Britain*. 2nd ed. Oxford, UK: Oxford University Press.

Google Developers. 2020. “Overview | Geolocation API.” *Google Developers*. Retrieved 3 August 2020. https://developers.google.com/maps/documentation/geolocation/overview.

Greenacre, Michael. 2017. *Correspondence Analysis in Practice*. Boca Raton, FL: CRC Press.

Guess, Andrew M., Pablo Barberá, Simon Munzert, and Jung H. Yang. 2021. “The Consequences of Online Partisan Media.” *Proceedings of the National Academy of Sciences* 118(14):e2013464118.

Hargittai, Eszter and Amanda Hinnant. 2008. “Digital Nequality: differences in Young Adults’ Use of the Internet.” *Communication Research* 35(5):602-21.

Hauser, Robert M. , and John R. Warren. 1997a. “Socioeconomic Indexes for Occupations: a Review, Update, and Critique.” *Sociological Methodology* 27(1):177-298.

Hinds, Joanne and Adam N. Joinson. 2018. “What Demographic Attributes Do Our Digital Footprints Reveal? A Systematic Review.” *PLoS One* 13(11):e0207112.

Hodas, Nathan O. and Kristina Lerman. 2012. “How Visibility and Divided Attention Constrain Social Contagion.” 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. Pp. 249-57.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. “Prediction and Explanation in Social Systems.” *Science (New York, N.Y.)* 355(6324):486-88.

Holt, Douglas B. 1998. “Does Cultural Capital Structure American Consumption?” *Journal of Consumer Research* 25(1):1-25.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” *Science* 353(6301):790-94.

Jiang, Yuqin, Zhenlong Li, and Xinyue Ye. 2019. “Understanding Demographic and Socioeconomic Biases of Geotagged Twitter Users at the County Level.” *Cartography and Geographic Information Science* 46(3):228-42.

Karami, A., M. Lundy, F. Webb, and Y. K. Dwivedi. 2020. “Twitter and Research: a Systematic Literature Review Through Text Mining.” *IEEE Access* 8:67698-717.

Katz-Gerro, Tally. 1999. “Cultural Consumption and Social Stratification: leisure Activities, Musical Tastes, and Social Location.” *Sociological Perspectives* 42(4):627-46.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” *Proceedings of the National Academy of Sciences* 110(15):5802-5.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. “The Geometry of Culture: analyzing the Meanings of Class Through Word Embeddings.” *American Sociological Review* 84(5):905-49.

Krieger, N., D. R. Williams, and N. E. Moss. 1997. “Measuring Social Class in US Public Health Research: concepts, Methodologies, and Guidelines.” *Annual Review of Public Health* 18(1):341-78.

Kwon, Eun S., Eunice Kim, Yongjun Sung, and Chan Y. Yoo. 2014. “Brand Followers.” *International Journal of Advertising* 33(4):657-80.

Lazer, David, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. “Meaningful Measures of Human Society in the Twenty-First Century.” *Nature* 595(7866):189-96.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915):721-23.

Leo, Yannick, Eric Fleury, J. Ignacio Alvarez-Hamelin, Carlos Sarraute, and Márton Karsai. 2016. "Socioeconomic Correlations and Stratification in Social-Communication Networks." *Journal of the Royal Society Interface* 13:125.

Leo, Yannick, Márton Karsai, Carlos Sarraute, and Eric Fleury. 2018. "Correlations and Dynamics of Consumption Patterns in Social-Economic Networks." *Social Network Analysis and Mining* 8(1):1-16.

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. "Social Media Fingerprints of Unemployment." *PLoS One* 10(5):e0128692.

Luo, Shaojun, Flaviano Morone, Carlos Sarraute, Matías Travizano, and Hernán A. Makse. 2017. "Inferring Personal Economic Status from Social Network Location." *Nature Communications* 8(1):1-7.

Maglio, Tony. 2016. "TV Show Viewers Ranked by Wealth, From "Modern Family" to "Empire"." *TheWrap*. Retrieved 2 May 2020. https://www.thewrap.com/richest-poorest-tv-shows-modern-family-empire/.

Maglio, Tony. 2018. "Summer 2018 TV Shows With the Richest and Poorest Viewers (Photos)." *TheWrap*. Retrieved 2 May 2020. https://www.thewrap.com/summer-2018-tv-shows-richest-poorest-viewers-photos/.

Marsden, Peter V. 1987. "Core Discussion Networks of Americans." *American Sociological Review* 52(1):122-31.

McCormick, Tyler H., Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma S. Spiro. 2017. "Using Twitter for Demographic and Social Science Research: tools for Data Collection and Processing." *Sociological Methods & Research* 46(3):390-421.

Mihelj, Sabina, Adrian Leguina, and John Downey. 2019. "Culture Is Digital: cultural Participation, Diversity and the Digital Divide." *New Media & Society* 21(7):1465-85.

Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos. 2004. "Does Education Improve Citizenship? Evidence from the United States and the United Kingdom." *Journal of Public Economics* 88(9):1667-95.

Moseley, Nathaniel, Cecilia O. Alm, and Manjeet Rege. 2014. "User-Annotated Microtext Data for Modeling and Analyzing Users' Sociolinguistic Characteristics and Age Grading." Pp. 1-6 in 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS).

Nam, Charles B. and Mary G. Powers. 1965. "Variations in Socioeconomic Structure by Race, Residence, and the Life Cycle." *American Sociological Review* 30(1): 97-103.

Nenadic, Oleg and Michael Greenacre. 2007. “Correspondence Analysis in R, with Two- and Three-Dimensional Graphics: the ca Package.” *Journal of Statistical Software* 20(1):1-13.

Norbutas, Lukas and Rense Corten. 2018. “Network Structure and Economic Prosperity in Municipalities: a Large-Scale Test of Social Capital Theory Using Social Media Data.” *Social Networks* 52:120-34.

Oakes, J. Michael and Kate Andrade. 2017. “The Measurement Of Socioeconomic Status.” Pp. 23-42 in *Methods in Social Epidemiology*, edited by J. M. Oakes and J. S. Kaufman. San Francisco, CA: Jossey-Bass & Pfeiffer Imprint, a Wiley brand.

ONS. 2020. “Standard Occupational Classification (SOC) - Office for National Statistics.” Retrieved 1 May 2020. https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc

Park, Patrick, Minsu Park, and Michael W. Macy. 2018. “Economic Correlates of Diversity and Inequality Online Social Networks.” *Academy of Management Proceedings* 2018(1):18881.

Peterson, Richard A. 1992. “Understanding Audience Segmentation: from Elite and Mass to Omnivore and Univore.” *Poetics* 21(4):243-58.

Peterson, Richard A. and Roger M. Kern. 1996. “Changing Highbrow Taste: from Snob to Omnivore.” *American Sociological Review* 61(5):900-7.

Preoţiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras. 2015. “An Analysis of the User Occupational Class Through Twitter Content.” Pp. 1754-64 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). Beijing, China: Association for Computational Linguistics.

Preoţiuc-Pietro, Daniel, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. “Studying User Income Through Language, Behaviour and Affect in Social Media.” *PLoS One* 10(9):e0138717.

Prieur, Annick and Mike Savage. 2013. “Emerging Forms of Cultural Capital.” *European Societies* 15(2):246-67.

Reeves, Aaron. 2019. “How Class Identities Shape Highbrow Consumption: a Cross-National Analysis of 30 European Countries and Regions.” *Poetics* 76:101361.

Rodríguez-Hernández, Carlos F., Eduardo Cascallar, and Eva Kyndt. 2020. “Socio-Economic Status and Academic Performance in Higher Education: a Systematic Review.” *Educational Research Review* 29:100305.

Rose, David, David J. Pevalin, and Karen O’Reilly. 2005. *The National Statistics Socio-Economic Classification: Origins, Development, and Use*. Basingstoke, Hampshire ; New York: Palgrave Macmillan.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole B. Carnegie, Ryan J. Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei L. Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15):8398-403.

Savage, Mike, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013a. "A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment." *Sociology* 47(2):219-50.

Schlenker, Barry R. and Beth A. Pontari. 2000. "The Strategic Control of Information: Impression Management and Self-Presentation in Daily Life." Pp. 199-232 in *Psychological Perspectives on Self and Identity*. Washington, DC, US: American Psychological Association.

Sirin, Selcuk R. 2005. "Socioeconomic Status and Academic Achievement: a Meta-Analytic Review of Research." *Review of Educational Research* 75(3): 417-53.

Sloan, Luke, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. "Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data." *PLoS One* 10(3):e0115545.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. "Limits of Predictability in Human Mobility." *Science* 327(5968): 1018-21.

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2020. “Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field.” *Social Science Computer Review* 38(5):503-16.

Szalai, Georg. 2010. “Cable Shows with the Wealthiest Viewers.” *The Hollywood Reporter*. Retrieved 4 August 2020. https://www.hollywoodreporter.com/news/cable-shows-wealthiest-viewers-25905

Taylor, Marshall A. and Dustin S. Stoltz. 2020. “Concept Class Analysis: a Method for Identifying Cultural Schemas in Texts.” *Sociological Science* 7:544-69.

Tucker, Joshua A., Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. SSRN Scholarly Paper. ID 3144139*. Rochester, NY: Social Science Research Network.

Twitter. 2020. “GET Friends/Ids.” Retrieved 2 May 2020. https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-friends-ids

Twitter. 2022. “Advanced Filtering for Geo Data.” *Twitter Developer Platform*. Retrieved 14 December 2022. https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data.

US Bureau of Labour Statistics. 2020. “May 2019 National Occupational Employment and Wage Estimates.” Retrieved 3 August 2020. https://www.bls.gov/oes/current/oes_nat.htm

van Dam, Alje, Mark Dekker, Ignacio Morales-Castilla, Miguel Á. Rodríguez, David Wichmann, and Mara Baudena. 2021. “Correspondence Analysis, Spectral Clustering and Graph Embedding: applications to Ecology and Economic Complexity.” *Scientific Reports* 11(1):8926.

van Deursen, Alexander J. A. M. and Jan AGM van Dijk. 2014. “The Digital Divide Shifts to Differences in Usage.” *New Media & Society* 16(3):507-26.

van Deursen, Alexander J. A. M. and Ellen J. Helsper. 2015. “The Third-Level Digital Divide: Who Benefits Most from Being Online?.” *Pp*. 29-52 in *Communication and Information Technologies Annual*. Vol. 10, Studies in Media and Communications. Emerald Group Publishing Limited.

Veblen, Thorstein. 2017. *The Theory of the Leisure Class*. Boca Raton: Routledge.

Volkova, Svitlana and Yoram Bachrach. 2015. “On Predicting Sociodemographic Traits and Emotions from Communications in Social Networks and Their Implications to Online Self-Disclosure.” *Cyberpsychology, Behavior, and Social Networking* 18(12):726-36.

Volkova, Svitlana, Yoram Bachrach, and Benjamin Van Durme. 2016. “Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter.” Pp. 36-43 in 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService).

Wagner, Claudia, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. "Measuring Algorithmically Infused Societies." *Nature* 595(7866):197-204.

Weininger, Elliot B. 2005. "Pierre Bourdieu on Social Class and Symbolic Violence." Pp. 116-65 in *Approaches to Class Analysis*, edited by E. O. Wright. Cambridge, UK: Cambridge University Press.

Werliin, Rune. 2020. "New Study: Instagram Climbs the Ladder, TikTok Has a Long Way to Go." *AudienceProject*. Retrieved 5 August 2021. https://www.audienceproject.com/blog/key-insights/new-study-instagram-climbs-the-ladder-tiktok-has-a-long-way-to-go/

Wikipedia. 2020. "List of Supermarket Chains in the United States." Wikipedia.

Wojcik, Stefan and Adam Hughes. 2019. "How Twitter Users Compare to the General Public." *Pew Research Center: Internet, Science & Tech*. Retrieved 20 July 2021. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

YouGov. 2018. "The Most Popular Speciality Retail Stores in America | Consumer | YouGov Ratings." Retrieved 2 May 2020. https://today.yougov.com/ratings/consumer/popularity/speciality-retail-stores/all

Youyou, Wu, Michal Kosinski, and David Stillwell. 2015. "Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans." *Proceedings of the National Academy of Sciences* 112(4):1036-40.

Yu, Jingyuan and Juan Muñoz-Justicia. 2020. "A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science." *Future Internet* 12(5):91.

Author Biographies

Yuanmo He is a PhD student in the Department of Methodology at the London School of Economics and Political Science. His PhD research examines how daily behaviors and social interactions reflect and reinforce socioeconomic inequality, using large-scale digital trace data and advanced computational methods. More broadly, he is interested in not only harnessing the potential of data science and AI to improve our understanding of society and humanity, but also evaluating the impacts of such technologies.

Milena Tsvetkova is an Assistant Professor of Computational Social Science at the Department of Methodology at the London School of Economics and Political Science. Her work uses large-scale online experiments, network analysis, machine learning, and computational modeling to study fundamental social phenomena such as cooperation, contagion, segregation, and inequality.