The London School of Economics and Political Science

Department of Methodology

Socioeconomic Inequality in Daily Behaviour and Social Interactions: Evidence from Digital Trace Data

Yuanmo He

A thesis submitted to the Department of Methodology of the London School of Economics and Political Science for the degree of Doctor of Philosophy

London

September 2024

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately **57,240** words (counting everything).

Statement of co-authored work

I confirm that Chapter 3 and 4 was jointly co-authored with Milena Tsvetkova, and I contributed 70% of the work.

Yuanmo He

贺渊默

Acknowledgements

> Before starting my PhD,
> looking at the statistics of PhD students suffering from mental health issues,
> I was worried.
> I was like, wow, 50 per cent of PhD students end up
> being happy?
> How is that going to help with my comedy career?!

That was a joke I used in my stand-up comedy sets. Jokes aside, I did suffer from mental health issues during my PhD, including a significant period of depression. In fact, since I entered my PhD during COVID-19 in September 2020, my whole four years have been shadowed by the lingering effects of the lockdowns and uncertainty on my physical and mental health. I was sure I would need an extension, but fortunately, it turned out that I could finish on time. Even now, as I finish up my PhD thesis, I am still in a suboptimal physical and mental state. I look forward to finishing my thesis and starting a new chapter of my life.

Nevertheless, I do not regret doing this PhD. Not at all. Despite all that, my PhD experience has been positive and rewarding. That is mainly owing to the amazing people I want to acknowledge here.

I want to thank my primary supervisor, Milena. You are the best supervisor I could have hoped for. Every supervision meeting with you makes me feel empowered and inspired. You provide the exact support and guidance I need with the right balance. You are straightforward, sharp, strict, rigorous but also kind, supportive, compassionate, and funny. You give me enough freedom to pursue my research interests with helpful guidance along the way. You give me enough push when I need motivation while offering sufficient support when I need to slow down. I really enjoy your supervision and our collaboration, and I hope we have more opportunities to collaborate in the future.

I want to thank everyone in the LSE Department of Methodology. Thanks to all of you, I spent the last five years of my Master's and PhD in the best learning and work environment. Thank you, Ken, for being a fantastic secondary supervisor and providing the guidance, resources, opportunities, and research freedom I need. Thank you, my fellow PhD students, past and present: Thiago, Christian, Poorvie, Noam, Denise, Oriol, Nancy, Akriti, Qi, Roni, Amal, Rosie, Jasmine, Katya, Izzi, Midanna, Ross, Chuyao, Wendy, Christy, and Islam. Thank you for making the office fun and lively. I will miss the banters, gossips, lunches, drinks and other activities we had. A special shout out to Oriol, my 'work husband'. You are the best friend I made from our office, and I have learned so much from you both as a person and researcher. I also want to thank our fantastic visiting students: Ana, Longsun, Simon, Livio, Johan, and Anatasia. We all had great times together. Thanks to my friends from my Masters, Pauline, Antonio, Natan, Davide, MJ, Ignacio, Alix, Parker, Silja, Kevin, Wenbo, Athena, Marie, Julia, Laila, Roy, Merle, Jinjie. You made my experience in the department amazing from the very beginning. Thanks to all the academic and professional services staff in the department. I won't list names here, but please know I am grateful to all of you. I do want to shout out to Camilya, because the PhD students all rely so much on you to make things work. You really take outstanding care of us.

I want to thank my closest friends, who have always been there for me. Thanks to my primary WeChat friend group members: Hu Yang, Ning Rui, Jia Ming, and Little Wang. Although we are located separately in the UK, China, and the US, we are so well connected. I would not survive this PhD without our group chats and video calls. Thanks to Haohao, the best friend I made during my undergraduate. Although I am more senior than you in age and academic journey, I am learning so much from you. Thanks to Zihao Tian, the best friend I made since I came to the UK. You encouraged me to be more outgoing and be myself. Thanks to Annie, the best friend I made in recent years. It is quite hard to have the deep connections we do, especially after becoming adults for a while. I am learning a lot from you, and thank you for being my primary contact during my period of depression. To all my closest friends, having any of you as a close friend is a privilege, so I am deeply grateful to have all of you in my life.

During my PhD, I had an unhealthy and narrow vision of focusing on my PhD without a life. I want to thank the groups outside my daily circle that give me a bit of life. Thanks to the friends I made in the Inequality Workshop at the International Centre for Theoretical Physics, the Complex Systems Summer School at the Santa Fe Institute, and the Beginner's Stand-up Comedy Course at Angel Comedy Club.

The making of a PhD does not start with graduate degrees. I want to thank all the professors who encouraged me to cultivate my research interests and pursue an academic career during my undergraduate degree. Particularly, thank you, Ozan, David, and Dylan. I also want to thank all the teachers who encouraged my curiosity in the past, even before university. Especially, my primary school Chinese Literature teacher, Wan Fangfang, and my secondary school Chinese Literature teacher, Shen Zhifang.

My most enormous gratitude goes to my parents. Thank you for supporting me in all possible ways to pursue my dreams. As a child, I thought our family was just ordinary, nothing special. But as I grew up, knew more people and experienced more of the world, I gradually realised the preciousness of our ordinary yet happy and harmonious family. Having a family like ours requires a lot of luck and work. I thank fate for the luck and you two for the work. Unfortunately, pursuing my dreams means staying so far away from you. Still, I am glad we have video calls every week. Thank you for teaching and cultivating, through your words and actions, all the valuable qualities I now possess. Thank you for your openness, tolerance, and rationality, which have allowed me to become a free, independent, and critical thinker. Thank you for your honesty and integrity, which have given me my own principles. Thank you for your love, compassion, and kindness, for making me feel that I always have something to fall back on, and for giving me the ability to care for others. It is a happy coincidence that this year marks your thirtieth year of marriage, and my thesis submission date is close to your anniversary date. Our family is not big on ceremonies, but please allow me to devote my PhD thesis to your marriage and our family. I love you, mom and dad. Our family will continue to be happy and harmonious no matter what happens.

Finally, I want to thank myself. You made it. Despite the love and support you get from your family and friends, there are things you must face alone. Only you know what you have been through. You made friends and mistakes. There are things you are proud of and things you regret. You have been through confusion, struggles, crises, heartbreaks, and depression. All those are helping you to grow and making you a better person. Please keep having a growth mindset; keep believing that everything that happens can be good for you. Most importantly, please remember to be yourself. Your biggest regrets are all from when you were trying to be someone else. There is nothing to regret if you just be your authentic self. At the end of the

seven-book series "Stories about Ming Dynasty ", after telling the stories of emperors, warlords, generals, and ministers of three hundred years, the author Dangnian Mingyue wrote the stories of Xu Xiake, an explorer who travelled throughout China, not driven by power, fame, nor money, but just interest. With that, he concludes:

"There is only one success—to live your life in your own way."

Abstract

This PhD thesis leverages large-scale digital trace data and advanced computational methods to examine how socioeconomic inequality is reflected and reinforced in daily life and social interactions. Grounded in Bourdieu's theory of economic, cultural, and social capital, the thesis comprises three empirical papers that explore different dimensions of socioeconomic inequality. The first paper proposes and validates a method to estimate individual Twitter users' SES based on the brands they follow. Rooted in Bourdieu's definition of socioeconomic status, the method measures a combination of economic and cultural capital. The SES estimates show significant correlations with traditional SES proxies, including income, education, and occupational social class. The second paper delves into the relationship between economic and cultural capital by utilising newly available mobile-tracking data to study inequality in daily consumption. Incorporating theories of conspicuous consumption, cultural omnivorousness, and inconspicuous consumption, the study presents a coherent theoretical framework suggesting that SES is positively associated with consumption diversity and offers large-scale empirical evidence supporting the hypothesis. The third paper utilises the SES estimates from the first paper to illustrate that Twitter users with higher SES tend to have higher social capital and more advantageous communication behaviour. It also shows that while high and low SES users mostly talk about similar topics, they tend to use different hashtags and have divergent sentiments towards immigration. Collectively, the thesis demonstrates the social and cultural factors in the persistence of inequality with large-scale digital trace data. The thesis not only extends existing social theories with innovative data and methods but also bridges the gap between theory-driven and data-driven research traditions.

Table of Contents

List of Figures

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 4

Chapter 5

Chapter 1 Introduction

One aspect of the complex nature of socioeconomic inequality is that it is manifested, maintained, and reproduced in daily behaviour and social interactions. People display their socioeconomic status (SES), compare SES with others, and experience socioeconomic inequality in daily life. Such relentless experience of the manifestation of socioeconomic inequality could contribute to the maintenance and reproduction of inequality (Bourdieu 1984; Kraus, Park, and Tan 2017). For example, in more unequal societies, people tend to pay more attention to and spend more money on status goods (Heffetz 2011; Walasek, Bhatia, and Brown 2018; Walasek and Brown 2015), which may make people work more (Bowles and Park 2005), save less (Wisman 2009), accrue more debt (Christen and Morgan 2005), or even declare bankruptcy (Perugini, Hölscher, and Collie 2016). Even in brief interactions, people use speech patterns and clothing as cues of socioeconomic status, which affects their judgement of the competence of others and could lead to unfavourable treatment of people of low SES on the job market (Kraus et al. 2019, 2017; Oh, Shafir, and Todorov 2020).

The recent rise of computational social science (CSS) offers unprecedented opportunities to study socioeconomic inequality in daily behaviour and social interactions. With the increasing availability of digital trace data, researchers can obtain data about daily behaviours and social interactions with a minimum obtrusion in real-time, at a low cost, and on a large scale. For example, researchers have used social media and mobile network data to link economic development to social capital at the level of geographical units. The research shows that individuals who live in areas with a high local development index tend to have more diverse networks with bridges that span greater geographic distances (Eagle, Macy, and Claxton 2010; Norbutas and Corten 2018). Others managed to obtain proprietary individual-level data to link SES to patterns in social interactions and consumption behaviours. It has been shown that people mostly communicate with their own or neighbouring SES groups, and apart from that, people communicate more with others from higher SES than lower SES (Leo et al. 2016). Also, SES has been found to be strongly associated with purchase patterns, and high SES tend to be correlated with more diverse purchases across product and service categories and merchants (Dong et al. 2020; Kalinin, Vaganov, and Bochenina 2020; Leo et al. 2018). Most recently, by linking a large sample of Facebook users in the U.S. to representative survey data, researchers find that economic connectedness (the proportion of high-SES friends among people with low SES) strongly correlates with upward income mobility (Chetty et al. 2022a). They also illustrate that variations of economic connectedness are explained mainly by different exposure to high-SES people and the tendency to befriend high-SES people conditional on exposure (Chetty et al. 2022b).

Building upon the existing CSS studies on socioeconomic inequality in daily behaviour and social interactions, my PhD thesis aims to study whether and how socioeconomic inequality is related to daily consumption, social capital, and communication behaviours. The thesis contributes to the CSS studies on socioeconomic inequality in daily behaviour and social interactions with the following objectives. First, this thesis develops a method to make it easier to conduct CSS research on socioeconomic inequality at the individual level. Second, this thesis introduces a direction of focus on consumption preferences and patterns, both to measure SES and study inequality. Third, the thesis advances the knowledge of how socioeconomic inequality is reflected and reinforced on social media platforms. Fourth, this thesis embeds

research more firmly in sociological theory, bridging newer interdisciplinary data-driven approaches and more traditional social science research.

The PhD thesis consists of three empirical papers to achieve the aims and objectives. The first paper, "A Method for Estimating Individual Socioeconomic Status of Twitter Users", is presented in Chapter 3. The first paper proposes a method to estimate the individual SES of Twitter users. Most of the existing CSS research on inequality is conducted at the level of geographical units or uses proprietary individual-level data. There is a constrained linkage between social media data and users' SES information and limited methods to estimate individual users' SES (Baghal et al. 2021; Ghazouani et al. 2019; Hinds and Joinson 2018; Stier et al. 2019). The first paper proposes a new approach to address the problem. Following Bourdieu (1984), the paper argues that the commercial and entertainment accounts that Twitter users follow reflect their economic and cultural capital. Hence, we can use the following to infer the users' SES. Adapting a political science method for inferring political ideology (Barberá et al. 2015), 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. The paper, co-authored with Dr Milena Tsvetkova, has already been published in *Sociological Methods & Research* (He and Tsvetkova 2023).

The second paper, "Omnivorous, Inconspicuous, and Niche: High Socioeconomic Status is Associated with Diverse Consumption", is presented in Chapter 4. The second paper digs further into the relationship between economic and cultural capital, utilising newly available mobile-tracking data to study inequality in daily consumption. Drawing on insights from sociology, social psychology, and consumer research, the paper integrates cultural omnivorousness and inconspicuous consumption theories and argues that high SES is associated with more diverse consumption practices. Consumption practices are determined by a combination of economic, social, and cultural forces. This bundling dictates that lower economic constraints leave more room to diversify consumption along cultural and social aspects in the form of omnivorous or lifestyle-based niche consumption. The paper analyses mobile tracking data of U.S. residents' visits to various stores to present evidence for the hypothesis. The results show that high SES, whether measured by income or education, is significantly associated with diverse consumption across brands and price levels. We further demonstrate that the associations cannot be explained by simple geographic constraints, including geographic mobility of the residents and local availability of the stores, so deeper social and cultural factors must be at play. The findings illustrate and quantify socioeconomic divisions in daily consumption practices, bearing further evidence for the pervasiveness and inevitability of socioeconomic inequality in daily life. The findings also provide further support to the underlying principle in the first paper, which is that consumption preferences can be used to predict SES.

The third paper, "Socioeconomic Inequality in Social Capital and Communication Behaviour on Twitter", is presented in Chapter 5. The third paper applies the method developed in the first paper to study how socioeconomic status is related to social capital and communication behaviour. On the one hand, the paper continues the recent efforts in quantifying the relationship between socioeconomic outcomes and social capital in digital communication

networks (Chetty et al. 2022a; Eagle et al. 2010; Luo et al. 2017; Norbutas and Corten 2018). The paper establishes that higher SES Twitter users have higher social capital across different measures of social capital. On the other hand, compared with the existing scattered evidence, this paper provides a more comprehensive picture of the relationship between SES and communication behaviour. The paper demonstrates that higher SES users use more complex and future-oriented language in their tweets. It also shows that while high and low SES users mostly talk about similar topics, they tend to use different hashtags and have divergent sentiments towards immigration. These findings reveal that socioeconomic inequalities are not only reflected but also potentially reinforced on social media, underscoring the critical roles of social capital and communication behaviour. The study highlights the need for further research to explore the underlying mechanisms and integrate SES as a critical factor in social media studies. The findings also further establish the utility of the method proposed in the first paper.

This PhD project places significant emphasis on Twitter due to its crucial role in society and its widespread use in academic research. Events and posts on Twitter have far-reaching impacts on societal outcomes in critical areas such as health, politics, and social movements (Murthy 2024). Twitter is one of the most popular social media platforms used for CSS research, with the number of Twitter-related studies steadily increasing (Karami et al. 2020). This prominence is not only due to its societal influence but also because of the platform's historically accessible data. Until early 2023, Twitter provided a well-developed Academic Track of its application programming interface (API) that was freely available to academic researchers. Unfortunately, the free Academic Track of the Twitter API has since been discontinued, significantly limiting future access to large-scale Twitter data. This change has already impacted the data collection process for the third paper in this dissertation, although not critically.

Nevertheless, the focus of this thesis on Twitter remains valuable. Before the API change, a key strength of this thesis' focus on Twitter was that the findings would be easily replicable and directly applicable to future Twitter research for many researchers. Now, this thesis holds unique value in maximising the use of large-scale Twitter data while accessible, especially in the relatively underexplored area of socioeconomic inequality on the platform. Although future research using Twitter may be more restricted, the findings from this thesis remain replicable and directly applicable to researchers with sufficient funding or alternative data access. Moreover, while this PhD research primarily focuses on Twitter due to its data availability at the time, the insights gained extend beyond this platform. The findings are relevant to broader social media research and could be applied to future studies of other platforms.

In the subsequent chapters of this PhD thesis, Chapter 2 begins with a literature review, outlining the key concepts and theoretical frameworks that underpin the research. Chapters 3, 4, and 5 present the three empirical studies that form the core of this thesis. Finally, Chapter 6 summarises the findings and discusses the overall conclusions drawn from this body of work.

References

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). doi: 10.1140/epjds/s13688-021-00286-7.

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. doi: 10.1177/0956797615594620.

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

Bowles, Samuel, and Yongjin Park. 2005. 'Emulation, Inequality, and Work Hours: Was Thorsten Veblen Right?' *The Economic Journal* 115(507):F397–412. doi: 10.1111/j.1468-0297.2005.01042.x.

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. 2022a. 'Social Capital I: Measurement and Associations with Economic Mobility'. *Nature* 608(7921):108–21. doi: 10.1038/s41586-022-04996-4.

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. 2022b. 'Social Capital II: Determinants of Economic Connectedness'. *Nature* 608(7921):122–34. doi: 10.1038/s41586-022-04997-3.

Christen, Markus, and Ruskin M. Morgan. 2005. 'Keeping Up With the Joneses: Analysing the Effect of Income Inequality on Consumer Borrowing'. *Quantitative Marketing and Economics* 3(2):145–73. doi: 10.1007/s11129-005-0351-1.

Dong, Xiaowen, Eaman Jahani, Alfredo J. Morales, Burçin Bozkaya, Bruno Lepri, and Alex ´Sandy' Pentland. 2020. 'Purchase Patterns, Socioeconomic Status, and Political Inclination'. *The World Bank Economic Review* 34(Supplement_1):S9–13. doi: 10.1093/wber/lhz008.

Eagle, N., M. Macy, and R. Claxton. 2010. 'Network Diversity and Economic Development'. *Science* 328(5981):1029–31. doi: 10.1126/science.1186605.

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.

He, Yuanmo, and Milena Tsvetkova. 2023. 'A Method for Estimating Individual Socioeconomic Status of Twitter Users'. *Sociological Methods & Research* 00491241231168665. doi: 10.1177/00491241231168665.

Heffetz, Ori. 2011. 'A Test of Conspicuous Consumption: Visibility and Income Elasticities'. *The Review of Economics and Statistics* 93(4):1101–17. doi: 10.1162/REST_a_00116.

Hinds, Joanne, and Adam N. Joinson. 2018. ‘What Demographic Attributes Do Our Digital Footprints Reveal? A Systematic Review’. *PLOS ONE* 13(11):e0207112. doi: 10.1371/journal.pone.0207112.

Kalinin, Alexander, Danila Vaganov, and Klavdiya Bochenina. 2020. ‘Discovering Patterns of Customer Financial Behavior Using Social Media Data’. *Social Network Analysis and Mining* 10(1):77. doi: 10.1007/s13278-020-00690-3.

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. doi: 10.1109/ACCESS.2020.2983656.

Kraus, Michael W., Jun Won Park, and Jacinth J. X. Tan. 2017. ‘Signs of Social Class: The Experience of Economic Inequality in Everyday Life’. *Perspectives on Psychological Science* 12(3):422–35. doi: 10.1177/1745691616673192.

Kraus, Michael W., Brittany Torrez, Jun Won Park, and Fariba Ghayebi. 2019. ‘Evidence for the Reproduction of Social Class in Brief Speech’. *Proceedings of the National Academy of Sciences* 116(46):22998–3. doi: 10.1073/pnas.1900500116.

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). doi: 10.1098/rsif.2016.0598.

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):9. doi: 10.1007/s13278-018-0486-1.

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. doi: 10.1038/ncomms15227.

Murthy, Dhiraj. 2024. ‘Sociology of Twitter/X: Trends, Challenges, and Future Research Directions’. doi: 10.1146/annurev-soc-031021-035658.

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. doi: 10.1016/j.socnet.2017.06.002.

Oh, DongWon, Eldar Shafir, and Alexander Todorov. 2020. ‘Economic Status Cues from Clothes Affect Perceived Competence from Faces’. *Nature Human Behaviour* 4(3):287–93. doi: 10.1038/s41562-019-0782-4.

Perugini, Cristiano, Jens Hölscher, and Simon Collie. 2016. ‘Inequality, Credit and Financial Crises’. *Cambridge Journal of Economics* 40(1):227–57. doi: 10.1093/cje/beu075.

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2019. ‘Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field’. *Social Science Computer Review* 0894439319843669. doi: 10.1177/0894439319843669.

Walasek, Lukasz, Sudeep Bhatia, and Gordon D. A. Brown. 2018. ‘Positional Goods and the Social Rank Hypothesis: Income Inequality Affects Online Chatter about High- and Low-Status Brands on Twitter’. *Journal of Consumer Psychology* 28(1):138–48. doi: 10.1002/jcpy.1012.

Walasek, Lukasz, and Gordon D. A. Brown. 2015. ‘Income Inequality and Status Seeking: Searching for Positional Goods in Unequal U.S. States’. *Psychological Science* 26(4):527–33. doi: 10.1177/0956797614567511.

Wisman, Jon D. 2009. ‘Household Saving, Class Identity, and Conspicuous Consumption’. *Journal of Economic Issues* 43(1):89–114.

Chapter 2 Literature Review

The definition and measurement of socioeconomic status.

This PhD project focuses on socioeconomic inequality, which I use as a simpler expression for the inequality of socioeconomic status. Thus, it is necessary to have a reasonable definition and measurement of socioeconomic status before studying its inequality. Socioeconomic status (SES) is a concept that describes an individual's social and economic position relative to others. 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 2019; Chan and Goldthorpe 2007; Flemmen, Jarness, and Rosenlund 2019; Hauser and Warren 1997a; 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 measuring 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 (EGP) class schema (Erikson and Goldthorpe 1992).

This PhD project adopts Bourdieu's (1984, 1986) definition of SES due to its substantial influence on both the area of social networks and the area of consumption. Bourdieu (1984, 1986) views individuals' 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. Bourdieu (1984) suggests that there is an overall volume of all forms of capital, but there are also separate dimensions for each form. Bourdieu (1986) emphasises the fungibility of these forms of capital and argues that the outcomes of possessing social or cultural capital can be ultimately reducible to economic capital. However, he believes that compared with the accumulation and exchange of economic capital, the dynamics of cultural and social capital and the exchange of these forms are less transparent and certain (Portes 1998). For Bourdieu, the issue of socioeconomic inequality is that advantaged individuals not only benefit from the possession of these forms of capital, but also can access and exchange them. Bourdieu (1984) uses the notion of *habitus* to describe the socially ingrained ways in which individuals perceive and act in the world, shaped by their economic, cultural, and social capital. Limited by their habitus, it is much harder for disadvantaged individuals to utilise all forms of capital to improve their overall socioeconomic status. Therefore, it is essential to consider all forms of capital when defining and measuring SES and analyse the dynamic relations between them when studying socioeconomic inequality.

Bourdieu's (1984, 1986) definition of SES is essential to the area of inequality and social capital. Credited as the first systematic contemporary analysis of social capital, Bourdieu's analysis of social capital is viewed as the key source for the studies of how social networks benefit individuals differently (Gelderblom 2018; Julien 2015; Lin 1999; Portes 1998). Social capital and social networks are two closely related concepts; many measures of social networks are viewed as forms of social capital (Jackson 2020; Lin 1999; Sabatini 2009). Bourdieu's emphasis on the fungibility of economic, cultural, and social capital provides a useful theoretical framework when studying inequality and social capital. Following Bourdieu's view of SES, different network measures may represent unequal access or possession of social capital. Due to the interchangeability of the different forms of capital, the unequal access or possession of social capital could eventually contribute to overall socioeconomic inequality.

Bourdieu's definition of SES is also influential in the area of consumption research. According to Bourdieu, consumption is not just determined by economic capital but by a combination of the different forms of capital. Bourdieu (1984) shows that in the 1970s in France, not only cultural consumption, such as literature and art, but also everyday consumption, such as clothing and eating, could be grouped by taste and correlated with SES in the form of unequal distribution of the different forms of capital. Bourdieu especially emphasises how cultural capital shapes taste, which then affects economic and cultural consumption. Cultural capital adds the cultural aspect to socioeconomic inequality and connects education, cultural industries, and stratification (Warde 2015). As cultural capital is obtained by upbringing or educational training, taste is more subtle and requires more effort to acquire than economic capital. Unequal cultural capital is reproduced effectively, as upbringing shapes cultural capital, which then affects educational attainment (Bourdieu and Passeron 2000; DiMaggio 1982).

As this PhD project will use Bourdieu's definition of SES, it is natural also to consider his measurement of SES. In his influential book *Distinction* (Bourdieu 1984), Bourdieu applied a dimensionality-reduction technique known as multiple correspondence analysis (MCA) on survey data from a sample of the French population in the 1960s containing income, occupation, and engagement in various cultural activities (e.g., reading, going to the concert, visiting museums). The technique allowed him to position individuals, occupations, and cultural activities on a two-dimensional graph. 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). He is then able to examine the relations between SES, economic capital, and cultural capital, as well as the role of habitus in the relations. It is worth noting that Bourdieu mainly focused on economic and cultural capital in *Distinction*. His analysis of social capital appears in his later work, but he did not provide similar measurement and statistical analysis for all three forms of capital.

The most influential attempt to incorporate all three forms of Bourdieu's capital comes from Savage et al. (2013). BBC's 2011 Great British Class Survey is a nationally representative survey of the UK population that contains questions targeting economic (income, savings, property value), cultural (engagement with cultural activities, e.g., going to concerts, visiting museums) and social (the number of social contacts and the occupations of these social contacts) capital. With data from the survey, Savage et al. (2013) separately compute economic, cultural and social capital for the participants and develop a seven-class schema of social class of the UK. Savage et al. (2013) suggest their schema is more informative than traditional occupation-based class schema due to the incorporation of different forms of capitals. However, Payne (2013) suggests that despite the conceptual differences, Savage et al.'s (2013) schema is very similar to the existing UK National Statistics Socioeconomic Classification (NS-SEC), which

is based on the occupation-based Erikson-Goldthorpe-Portocarero (EGP) class schema (Erikson and Goldthorpe 1992; Rose, Pevalin, and O'Reilly 2005).

The first and third empirical papers of this PhD project use a combination of economic and cultural capital as a proxy for SES and explore social capital as a dependent variable, while the second paper also examines the relationship between economic and cultural capital. The first paper of this PhD project (Chapter 3) develops a measure of SES similar to Bourdieu's (1984) measure of the overall volume of economic and cultural capital suitable for Twitter. To validate the developed measure, the paper compares it with commonly used indicators of SES, including income, education, and occupation. The third paper (Chapter 5) examines the socioeconomic inequality in social capital and communication behaviours on Twitter. However, before using a measure that combines economic and cultural capital and linking it to social capital, it is helpful to learn more about the interactions between economic and cultural capital. Daily consumption behaviours provide a suitable case for such purpose, and the second paper (Chapter 4) of the PhD project studies precisely this.

SES, cultural capital, and consumption.

Veblen's (1899) theory of conspicuous consumption is one of the earliest and most influential theories describing the relationship between consumption and SES. The theory is introduced in Veblen's book *Theory of the Leisure Class*. According to Veblen, the economic and technological development of a society contributes to the evolution of a leisure class, whose members enjoy the surplus produced by the working class and do not need to work as much (Trigg 2001). The effective operation of surplus leads to the acquisition and accumulation of private property. Thus, the accumulation of property becomes an indicator of one's competence, which grants status and honour in the social hierarchy. As the association between property and status becomes widely accepted, the display of wealth evolves into a useful way of establishing status. Veblen (1899) suggests that the key to displaying wealth is showing the capacity to afford waste. He uses the notion of *conspicuous leisure* to describe the practice of displaying wealth through engaging in extensive leisure activities that basically waste effort and time, and the concept of *conspicuous consumption* to represent the practice of displaying wealth through purchasing and using goods that often cost more than their practical value. Moreover, Veblen (1899) argued that as mobility in society increases, it is harder to keep people informed about their leisure activities, which makes conspicuous consumption a more effective display of wealth than conspicuous leisure.

Veblen (1899) regards conspicuous consumption as the most important determinant of consumer behaviour. He believes that as people of high social status display their wealth and status through conspicuous consumption, they also set up an ideal for people at the lower level of the social hierarchy. Because people aspire to obtain status by consumption, people at each level of the social hierarchy emulate the consumption of people at a higher level, referred to by Veblen as *pecuniary emulation*. This process may never end. As people of lower status catch up, people of higher status must find new goods of conspicuous consumption to distinguish themselves, creating new rounds of pecuniary emulation for people of lower status.

The concept of conspicuous consumption has proven useful in many cases. For example, it has been used in studies to explain the choice of cosmetic brands (Chao and Schor 1998), consumption of niche products (Schaefers 2014), and racial differences in the proportion of expenditure spent on visible goods (Charles, Hurst, and Roussanov 2009). Conspicuous

consumption also can be linked to the reproduction of inequality. Although conspicuous consumption has some positive effects, such as the perception of elevated status and short-term happiness, it compromises the budget for basic needs and long-term progress and could lead to longer work hours (Bowles and Park 2005; Kumar et al. 2021; Srivastava, Mukherjee, and Jebarajakirthy 2020).

However, researchers must be cautious about context when applying the concept of conspicuous consumption. The utility and relevance of the idea have been fluctuating depending on the historical and societal context (Patsiaouras and Fitchett 2012; Trigg 2001). Veblen coined the notion of conspicuous consumption during the Gilded Age, an era of rapid economic growth in the United States, and thus, an era of surplus and competition for status. However, soon after that, there were the two World Wars and the Great Recession between them. During these difficult times, public attitudes towards conspicuous consumption changed, and charitable activities became a more important way for wealthy people to attain social status (Galbraith 1989; Patsiaouras and Fitchett 2012). After the Second World War, as the economy recovered and post-industrial capitalism flourished in affluent Western societies, the relevance of conspicuous consumption revived and increased. Economic prosperity led to high levels of social mobility and, thus, a high demand for status-seeking and signalling. The prevalence of television and TV advertising significantly changed the delivery of conspicuous consumption practices and made the association between goods and status more visible (Galbraith 1989; Packard 1957).

Coming to the 21st century, two important challenges have been raised about the utility of conspicuous consumption. One challenge argues that contemporary capitalism breeds individualistic lifestyles, tastes, and identities instead of social class or status, so the relationship between socioeconomic status and consumption is fading (Mason 1998; McIntyre 1992). Another challenge is that status signalling now shifts to the consumption of inconspicuousness, non-ownership (experiential consumption), and authenticity. The association between status and consumption is no longer as strict and straightforward as before, and many of the traditional luxury brands that take the function of conspicuous consumption have changed their marketing strategy to aim for mass-market proliferation (Berger and Ward 2010; Eckhardt and Bardhi 2020; Eckhardt, Belk, and Wilson 2015).

Bourdieu's theory of taste and cultural capital provides a more powerful and exclusive process of status distinction than conspicuous consumption (Trigg 2001). The status distinction is achieved through taste, not just by confirming one's own status group's taste as superior but also by rejecting other groups' tastes as inferior. This negativity aspect increases the exclusivity of taste. Taste also provides a reasonable explanation for the shifting status signals of consumption. As conspicuous consumption becomes more affordable, people in higher SES need to reject the popular taste of consumption and develop new tastes, which manifests as the appreciation of inconspicuous, experiential, or authenticity consumption. Moreover, Bourdieu's theory of capital makes it possible to map lifestyles into social groups associated with different levels of capital in different forms. For example, as people with high cultural capital abandon conspicuous consumption, it is reasonable to expect that people who still practice conspicuous consumption would consist mainly of people with high economic but low cultural capital, and some status seekers with middle or low economic and low cultural capital. While people with high cultural and economic capital may pursue expensive, inconspicuous consumption, people with high cultural but low economic capital may practice the inexpensive options.

An alternative to Bourdieu's (1984) theory of cultural capital on cultural consumption is the cultural omnivorousness theory, which maintains that instead of the stratified highbrow-lowbrow taste, people in higher SES tend to enjoy more diverse cultural products (Peterson 1992). In the second paper (Chapter 4), I incorporate the existing theories of consumption inequality and propose to combine the recent theories of cultural omnivorousness and inconspicuous consumption. The second paper further examines the relationship between economic and cultural capital. It supports some assumptions made in the first paper about the effects of economic and cultural capital on consumption preferences. After the first two papers, we are better equipped in terms of theory, data, and method to further explore socioeconomic inequality in social capital and communication behaviours in the third paper.

SES and social capital

Social capital is a concept that captures the value embedded within individuals' social networks, often determining access to resources, information, and opportunities. However, social capital operates at multiple levels and is measured through various dimensions, leading to inconsistent findings regarding its relationship with socioeconomic outcomes (Portes 1998; Westlund and Adam 2010). A useful theoretical distinction of social capital, particularly for the scope of this PhD thesis, is between competitive and cooperative social capital (Gelderblom 2018; Julien 2015; Lin 1999; Portes 1998). This distinction allows a more nuanced understanding of how social capital might interact with socioeconomic status (SES) in digital and physical environments.

Competitive social capital refers to the form of social capital that provides individuals with competitive advantages through the size, strength, and structure of their social networks. Bourdieu's (1984, 1986) analysis of social capital mainly focused on competitive social capital, emphasising the instrumental role of social capital in providing individuals with resources embedded in social networks with expected returns. Although he does not use the term social capital, Granovetter's (1973) "strength of weak ties" is viewed as a key example of competitive social capital (Lin 1999; Portes 1998). Granovetter (1973) uses weak ties to represent the ties outside the closest associates and suggests that weak ties are influential as they provide extra information in contexts such as employment. Building on Granovetter, Burt (1992) develops the notion of "structural holes", representing the lack of closure in individuals' networks. Structural holes put individuals in favourable network positions for new information. Competitive social capital may also be found in the level of communities; Putnam's (2000) idea of bridging capital suggests that communities that are well-connected to other communities may enjoy competitive advantages over isolated communities.

Conversely, cooperative social capital describes the kind of social capital that people enjoy in a tightly connected community, where a high level of trust facilitates activities beneficial to everyone. Coleman (1988) acknowledges the role of networks as information channels for individuals but emphasises the benefits of community norms and sanctions for orienting individual actions toward collective interests. Similarly, Putnam (2000) approaches social capital as the norms and trust in a community that contribute to cooperation towards the common good. Putnam's (2000) notion of bonding capital suggests that people in densely connected communities may benefit from a sense of belonging and emotional support within the communities. Operationally, competitive social capital is better analysed at the individual level but can be aggregated to the group level. On the contrary, cooperative social capital is typically examined at the community level and is hard to disaggregate to the individual level.

As this PhD project examines socioeconomic inequality at the individual level, the emphasis naturally falls on competitive social capital. Nevertheless, it is useful to consider the theoretical distinction when interpreting relevant literature on the relationship between SES and social capital. The relationship between socioeconomic inequality and social capital is well-established in the literature (DiMaggio and Garip 2012; Granovetter 2005; Jackson 2021; Redhead and Power 2022). However, relevant research on large digital communication networks is constrained due to the limited availability of socioeconomic data. Only a few recent studies are available using proprietary data. The third paper (Chapter 6) in this thesis contributes to this emerging research by examining the socioeconomic inequality in social capital on Twitter. The paper also examines the socioeconomic inequality in communication behaviours on Twitter, an area where the literature remains fragmented and underdeveloped.

References

Berger, Jonah, and Morgan Ward. 2010. 'Subtle Signals of Inconspicuous Consumption'. *Journal of Consumer Research* 37(4):555–69. doi: 10.1086/655445.

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

Bourdieu, Pierre. 1986. 'The Forms of Capital'. Pp. 241–58 in *Handbook of Theory and Research for the Sociology of Education*, edited by J. Richardson. New York: Greenwood.

Bowles, Samuel, and Yongjin Park. 2005. 'Emulation, Inequality, and Work Hours: Was Thorsten Veblen Right?' *The Economic Journal* 115(507):F397–412. doi: 10.1111/j.1468-0297.2005.01042.x.

Burt, Ronald S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, Mass: Harvard University Press.

Chan, Tak Wing. 2019. 'Understanding Social Status: A Reply to Flemmen, Jarness and Rosenlund'. *The British Journal of Sociology* 70(3):867–81. doi: https://doi.org/10.1111/1468-4446.12628.

Chan, Tak Wing, and John H. Goldthorpe. 2007. 'Social Status and Newspaper Readership'. *American Journal of Sociology* 112(4):1095–1134. doi: 10.1086/508792.

Chao, Angela, and Juliet B. Schor. 1998. 'Empirical Tests of Status Consumption: Evidence from Women's Cosmetics'. *Journal of Economic Psychology* 19(1):107–31. doi: 10.1016/S0167-4870(97)00038-X.

Charles, Kerwin Kofi, Erik Hurst, and Nikolai Roussanov. 2009. 'Conspicuous Consumption and Race*'. *The Quarterly Journal of Economics* 124(2):425–67. doi: 10.1162/qjec.2009.124.2.425.

DiMaggio, Paul, and Filiz Garip. 2012. 'Network Effects and Social Inequality'. *Annual Review of Sociology* 38(1):93–118. doi: 10.1146/annurev.soc.012809.102545.

Duncan, Otis Dudley. 1961. ‘A Socioeconomic Index for All Occupations’. Pp. 109–38 in *Occupations and Social Status*. New York: Free Press.

Eckhardt, Giana M., and Fleura Bardhi. 2020. ‘New Dynamics of Social Status and Distinction’. *Marketing Theory* 20(1):85–102. doi: 10.1177/1470593119856650.

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. doi: 10.1080/0267257X.2014.989890.

Erikson, Robert, and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford [England] : New York: Clarendon Press ; Oxford University Press.

Flemmen, Magne Paalgard, 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. doi: https://doi.org/10.1111/1468-4446.12508.

Galbraith, John Kenneth. 1989. *A History of Economics: The Past as the Present*. Harmondsworth: Penguin Books.

Gelderblom, Derik. 2018. ‘The Limits to Bridging Social Capital: Power, Social Context and the Theory of Robert Putnam’. *The Sociological Review* 66(6):1309–24. doi: 10.1177/0038026118765360.

Granovetter, Mark. 2005. ‘The Impact of Social Structure on Economic Outcomes’. 18.

Granovetter, Mark S. 1973. ‘The Strength of Weak Ties’. *American Journal of Sociology* 78(6):1360–80. doi: 10.1086/225469.

Hauser, Robert M., and John Robert Warren. 1997a. ‘Socioeconomic Indexes for Occupations: A Review, Update, and Critique’. *Sociological Methodology* 27:177–298.

Hauser, Robert M., and John Robert Warren. 1997b. ‘Socioeconomic Indexes for Occupations: A Review, Update, and Critique’. *Sociological Methodology* 27(1):177–298. doi: 10.1111/1467-9531.271028.

Jackson, Matthew O. 2020. ‘A Typology of Social Capital and Associated Network Measures’. *Social Choice and Welfare* 54(2–3):311–36. doi: 10.1007/s00355-019-01189-3.

Jackson, Matthew O. 2021. ‘Inequality’s Economic and Social Roots: The Role of Social Networks and Homophily’.

Julien, Chris. 2015. ‘Bourdieu, Social Capital and Online Interaction’. *Sociology* 49(2):356–73. doi: 10.1177/0038038514535862.

Kumar, Bipul, Richard P. Bagozzi, Ajay K. Manrai, and Lalita A. Manrai. 2021. ‘Conspicuous Consumption: A Meta-Analytic Review of Its Antecedents,

Consequences, and Moderators'. *Journal of Retailing*. doi: 10.1016/j.jretai.2021.10.003.

Lin, Nan. 1999. 'Building a Network Theory of Social Capital''. *Connections* 22(1):28–51.

Mason, Roger S. 1998. *The Economics of Conspicuous Consumption: Theory and Thought since 1700*. Cheltenham, UK ; Northampton, MA: Edward Elgar.

McIntyre, Richard. 1992. 'Consumption in Contemporary Capitalism: Beyond Marx and Veblen'. *Review of Social Economy* 50(1):40–60.

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. doi: 10.2307/2091776.

Packard, Vance. 1957. *The Hidden Persuaders*. New York: McKay.

Patsiaouras, Georgios, and James A. Fitchett. 2012. 'The Evolution of Conspicuous Consumption' edited by B. Wooliscroft. *Journal of Historical Research in Marketing* 4(1):154–76. doi: 10.1108/17557501211195109.

Payne, Geoff. 2013. 'Models of Contemporary Social Class: The Great British Class Survey'. *Methodological Innovations Online* 8(1):3–17. doi: 10.4256/mio.2013.001.

Peterson, Richard A. 1992. 'Understanding Audience Segmentation: From Elite and Mass to Omnivore and Univore'. *Poetics* 21(4):243–58. doi: 10.1016/0304-422X(92)90008-Q.

Portes, Alejandro. 1998. 'Social Capital: Its Origins and Applications in Modern Sociology'. *Annual Review of Sociology* 24(1):1–24. doi: 10.1146/annurev.soc.24.1.1.

Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Redhead, Daniel, and Eleanor A. Power. 2022. 'Social Hierarchies and Social Networks in Humans'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1845):20200440. doi: 10.1098/rstb.2020.0440.

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.

Sabatini, Fabio. 2009. 'Social Capital as Social Networks: A New Framework for Measurement and an Empirical Analysis of Its Determinants and Consequences'. *The Journal of Socio-Economics* 38(3):429–42. doi: 10.1016/j.socec.2008.06.001.

Savage, Mike, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013. 'A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment'. *Sociology* 47(2):219–50. doi: 10.1177/0038038513481128.

Schaefers, Tobias. 2014. ‘Standing out from the Crowd: Niche Product Choice as a Form of Conspicuous Consumption’. *European Journal of Marketing* 48(9/10):1805–27. doi: 10.1108/EJM-03-2013-0121.

Srivastava, Abhinav, Srabanti Mukherjee, and Charles Jebarajakirthy. 2020. ‘Aspirational Consumption at the Bottom of Pyramid: A Review of Literature and Future Research Directions’. *Journal of Business Research* 110:246–59. doi: 10.1016/j.jbusres.2019.12.045.

Trigg, Andrew B. 2001. ‘Veblen, Bourdieu, and Conspicuous Consumption’. *Journal of Economic Issues* 35(1):99–115. doi: 10.1080/00213624.2001.11506342.

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.

Westlund, Hans, and Frane Adam. 2010. ‘Social Capital and Economic Performance: A Meta-Analysis of 65 Studies’. *European Planning Studies* 18(6):893–919. doi: 10.1080/09654311003701431.

Chapter 3 A Method for Estimating Individual Socioeconomic Status of Twitter Users

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 API, 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 non-significant correlations with other demographic variables. The proposed method opens new opportunities for innovative social research on inequality on Twitter and similar online platforms.

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 2007; 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 analysed 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 behaviour, 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 behaviours, 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, TikTok).

Nevertheless, it is hard to infer Twitter users' socioeconomic status. Twitter does not have a designated field that requires socioeconomic information. Some Twitter users state their occupations 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 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 2019; Chan and Goldthorpe 2007; 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 which 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 (EGP) 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 labelled 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 towards 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% 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 non-parametric 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 correspondence analysis, 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 neighbourhood 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 computational social science (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 correspondence analysis 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 towards 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 for social distinction. Veblen ([1899] 2017) made a similar argument about distinction but instead emphasised 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% 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 don't 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 correspondence analysis (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, 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 correspondence analysis (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 correspondence analysis (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 Multiple Correspondence Analysis (MCA) conducted by Bourdieu himself (Bourdieu 1984; Flemmen et al. 2018).

CA is a multivariate method to summarise and visualise 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 standardised 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 $c_j = \sum_{i=1}^{I} P_{ij}$, 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 $\mathbf{G_c} = \mathbf{D_c}\mathbf{V}$ for columns. As the original data matrix **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 $\mathbf{n}' = \frac{\mathbf{n}}{\sum_{i=1}^{I} n_i}$ and compute $\mathbf{g} = \mathbf{n}'^{T}\mathbf{G}_{\mathbf{r}}$. 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 University Ethical Review Board and the complete list of brands and their Twitter accounts required to replicate the results is available in the Supplementary 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' audience (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 & department stores, clothing & speciality retailers, chain restaurants, newspapers & 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 x 303 matrix (in comparison, the full matrix is 3,482,657 x 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 towards 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).

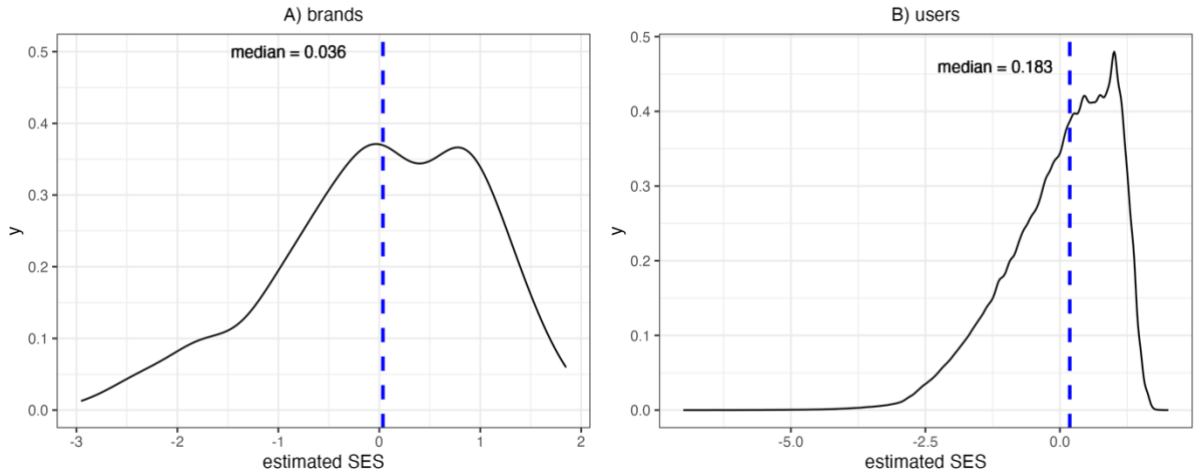

Figure 1. Density plot of the estimated SES for A) 339 brands and B) 3,482,657 users who follow them on Twitter.

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, 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.

We note that since SES is a composite concept, and our measure is operationalized to capture this multi-facetedness, 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 Supplementary 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*, *Toys "R" Us*, etc. 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 2010 was reported to have 48% of its viewers with 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 (2007), which shows that national newspapers tend to be read by people with higher social status.

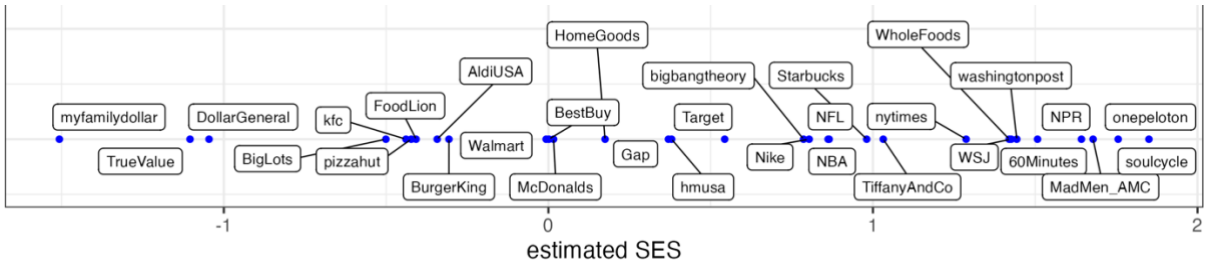

Figure 2. Estimated SES for a selected group of popular brands.

In a 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, 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, 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 $\rho = -0.464$, $p < 0.001$), while panel C) shows a positive association between the estimated SES and the proportion who hold a Master's or higher degree ($\rho = 0.444$, $p < 0.001$). 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 ($\rho = 0.320$, $p < 0.001$).

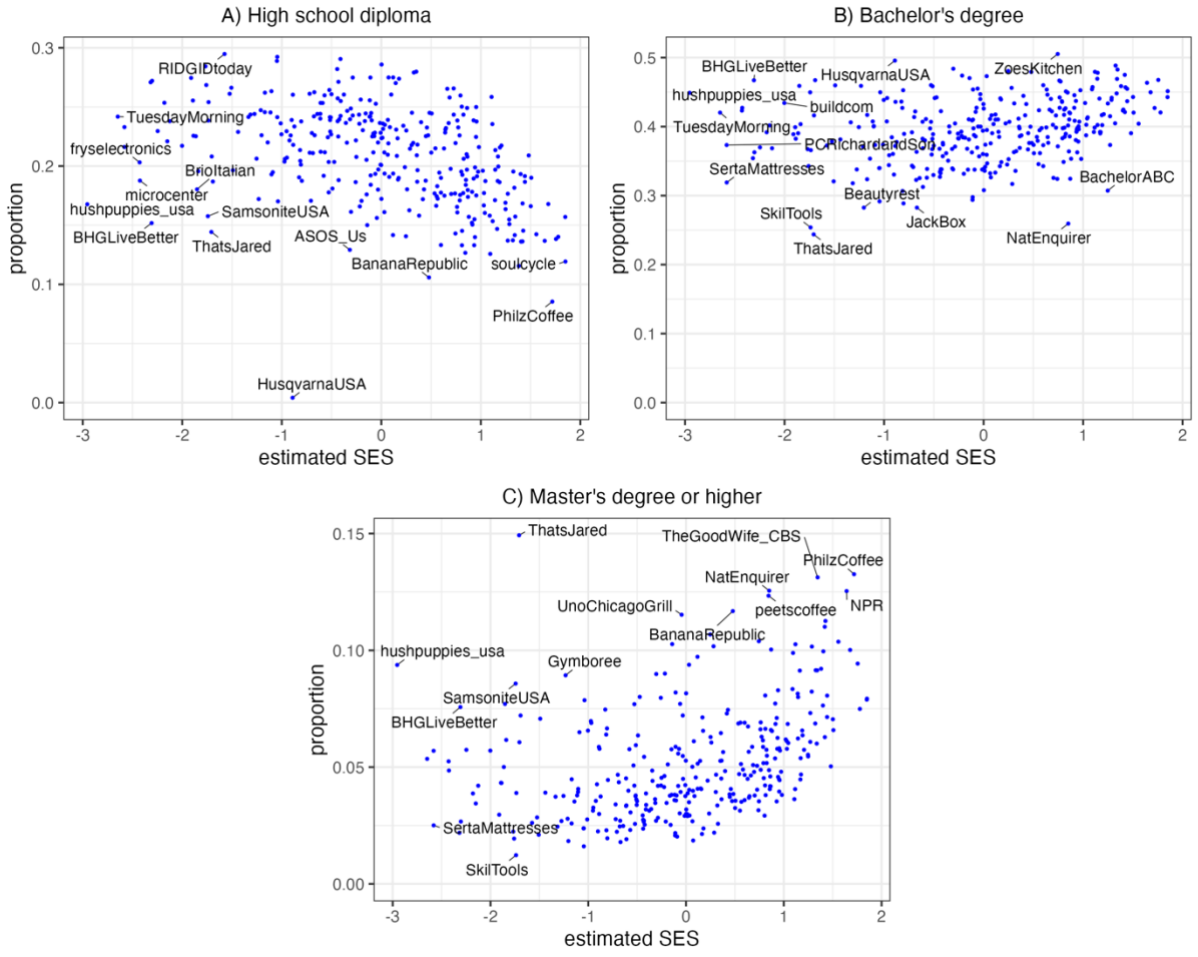

Figure 3. Relation between the educational composition of the brands' Facebook audience, as measured by the proportion who have earned at most the respective accreditation, and the brands' estimated SES.

The patterns in the plots and the correlation statistics show that the brands with higher estimated SES tend to have significantly smaller audience of at-most high school graduates, significantly larger audience with Master's or higher degrees, and somewhat larger audience of Bachelor degree holders. The latter represent 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 have 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 minimise the number of wrongly labelled titles, we include an additional filter: we manually inspect ten randomly sampled descriptions for each job title to identify text structures that contribute to mislabelling 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. Supplementary 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 colours 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 the Spearman's rank correlation between the median estimated SES and mean annual salary is 0.673 ($p < 0.001$). The Spearman's rank correlation between the median estimated SES and occupational class is -0.640 ($p < 0.001$). As a reference, in our sample, the Spearman's correlation between mean annual salary and class is -0.840 ($p < 0.001$). 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.

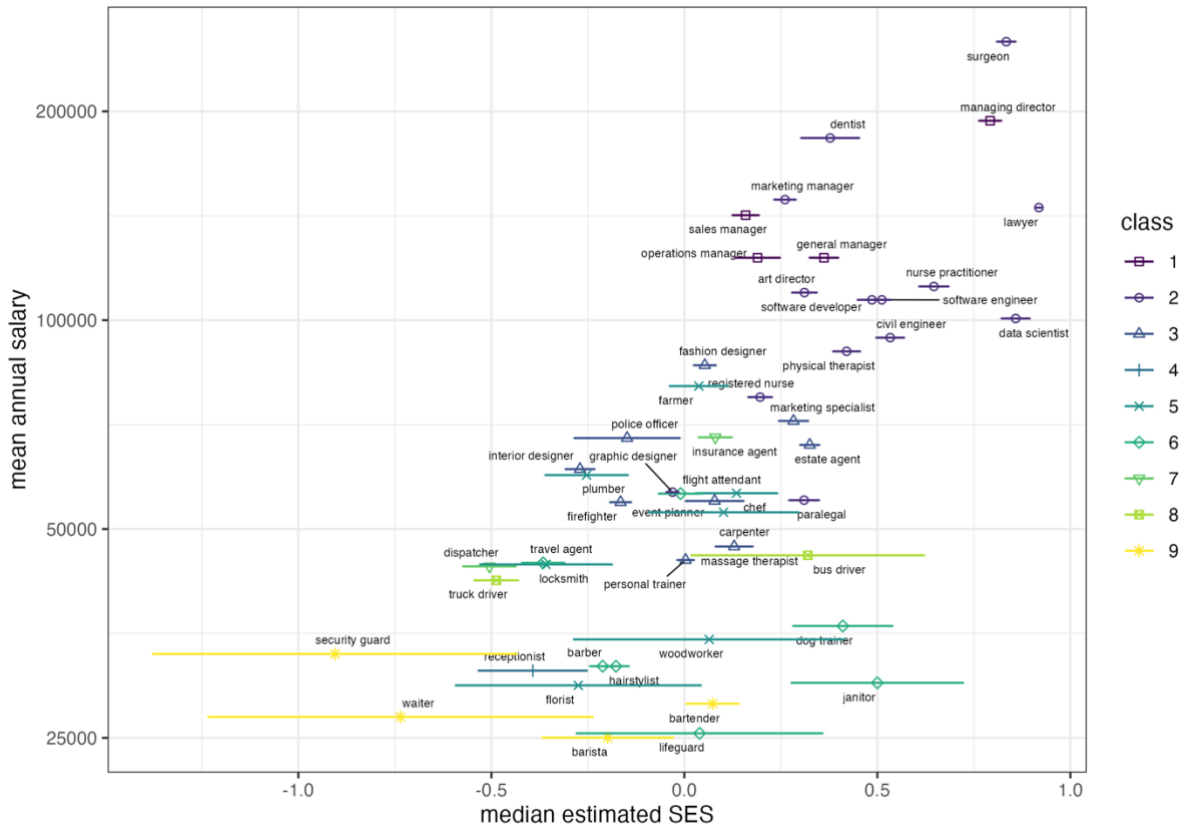

Figure 4. Relation between median estimated SES, mean salary, and occupational class for a set of 50 common job titles, estimated over 42,099 Twitter users who mention one of the titles in their profile description.

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 the 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 level data 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 < 0.001$) and the one between estimated SES and income level to be 0.188 ($p < 0.05$). As a reference, the Spearman correlation between income and education in the sample is 0.455 ($p < 0.001$), 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, $p < 0.001$, $N = 147$ in the case of two accounts; 0.344, $p < 0.001$, $N = 111$ for three accounts) but weaken for income (0.137, $p = 0.117$, $N = 131$ for two accounts; 0.156, $p = 0.117$, $N = 102$ 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 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.

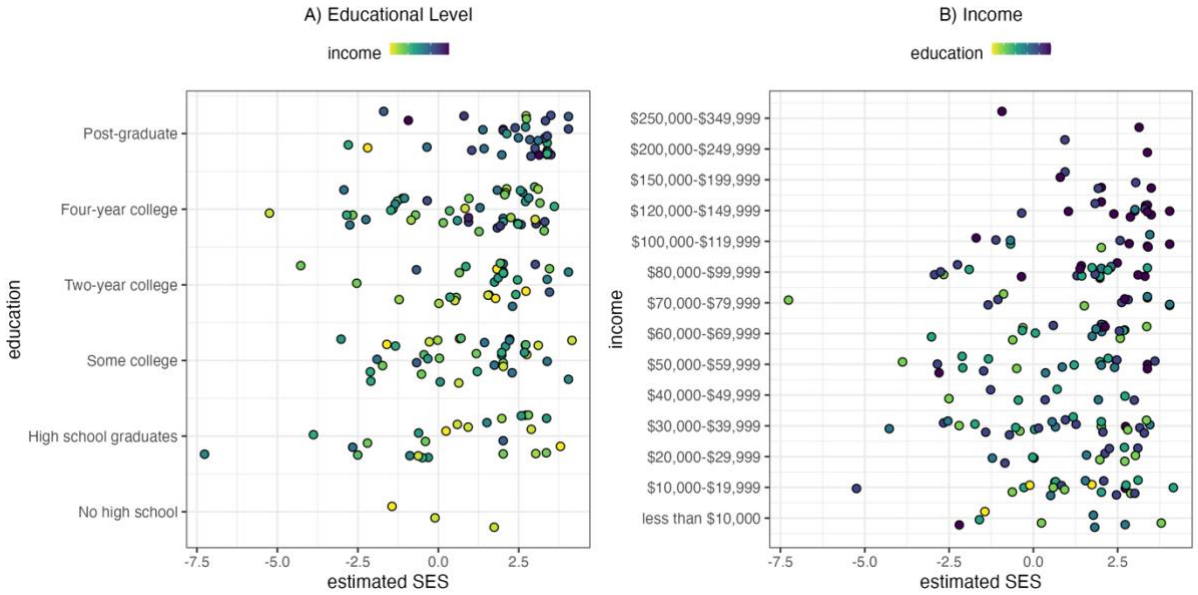

Figure 5. Relation between estimated SES and A) educational level and B) income for 200 (182 for B) survey respondents who follow at least one of the 339 brands on Twitter. The y-axis values are plotted with noise to improve visibility.

Divergent validity

So far, we focused on convergent validity, utilising 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 analyse 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 ($\rho = 0.114$, $p = 0.050$) and not associated with the proportion of male ($\rho = 0.034$, $p = 0.558$) nor female ($\rho = -0.037$, $p = 0.532$) users. These results suggest our SES measure for the brands is not capturing urban/rural nor 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 ($p < 0.01$) for age 18-24, 0.199 ($p < 0.001$) for 25-34, and 0.136 ($p < 0.05$) for 35-44. Conversely, the estimates have weak negative associations with older age groups: Spearman correlation coefficients of -0.135 ($p < 0.05$) for age 45-54, -0.224 ($p < 0.001$) for 55-64, and -0.117 ($p < 0.05$) 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 ($p < 0.001$), 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 ($p = 0.139$) and the t-test between male and female is 0.741 ($p = 0.460$). Similarly, there is no significant difference in estimated SES between the four racial groups (White, Black, Hispanic, Asian/other) categorised in the survey, regardless of whether we use an analysis of variance test ($p = 0.871$) 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 ($p = 0.279$). 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

[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 Supplementary 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.

Supplementary Table 5). Further, regression analyses, presented in Supplementary Table 6, show that controlling for age, gender, political ideology and race, there are still significant correlations between estimated SES and education ($p < 0.001$) and between estimated SES and income ($p < 0.05$).

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 behavioural 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 with 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 finetuning 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. Aggerated 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, e.g., 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 towards 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 achieves 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. 2019) but the resulting linked data could contribute in multiple ways. First, the data will allow re-validating 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 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 a 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 correspondence analysis 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 & Goldthorpe, 2007; 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 on 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 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 2021). 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 behaviours, 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 behaviours, opinions, and knowledge expressed on those platforms.

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.

Authors' Note

Data, codes, 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).

References

Abitbol, Jacob, Eric Fleury, and Márton Karsai. 2019. 'Optimal Proxy Selection for Socioeconomic Status Inference on Twitter'. *Complexity* 2019:e6059673. doi: 10.1155/2019/6059673.

Abitbol, Jacob Levy, and Márton Karsai. 2020. 'Interpretable Socioeconomic Status Inference from Aerial Imagery through Urban Patterns'. *Nature Machine Intelligence* 2(11):684–92. doi: 10.1038/s42256-020-00243-5.

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. doi: 10.1037/0003-066X.49.1.15.

Airoldi, Massimo. 2021. 'The Techno-Social Reproduction of Taste Boundaries on Digital Platforms: The Case of Music on YouTube'. *Poetics* 89:101563. doi: 10.1016/j.poetic.2021.101563.

Alderson, Arthur S., Azamat Junisbai, and Isaac Heacock. 2007. 'Social Status and Cultural Consumption in the United States'. *Poetics* 35(2):191–212. doi: 10.1016/j.poetic.2007.03.005.

Aletras, Nikolaos, and Benjamin Paul 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. doi: 10.1037/a0038781.

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 Mohammady, 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.

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. doi: 10.1073/pnas.1804840115.

Barberá, Pablo. 2015. 'Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data'. *Political Analysis* 23(1):76–91. doi: 10.1093/pan/mpu011.

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. doi: 10.1177/0956797615594620.

Berger, Jonah, and Morgan Ward. 2010. 'Subtle Signals of Inconspicuous Consumption'. *Journal of Consumer Research* 37(4):555–69. doi: 10.1086/655445.

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. doi: 10.1017/S0003055414000525.

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

Brady, Henry E., Sidney Verba, and Kay Lehman Schlozman. 1995. 'Beyond SES: A Resource Model of Political Participation'. *American Political Science Review* 89(2):271–94. doi: 10.2307/2082425.

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. doi: 10.1016/S0378-8733(86)80017-X.

Chan, Tak Wing. 2019a. 'Understanding Cultural Omnivores: Social and Political Attitudes'. *The British Journal of Sociology* 70(3):784–806. doi: 10.1111/1468-4446.12613.

Chan, Tak Wing. 2019b. 'Understanding Social Status: A Reply to Flemmen, Jarness and Rosenlund'. *The British Journal of Sociology* 70(3):867–81. doi: https://doi.org/10.1111/1468-4446.12628.

Chan, Tak Wing, and John H. Goldthorpe. 2007a. 'Class and Status: The Conceptual Distinction and Its Empirical Relevance'. *American Sociological Review* 72(4):512–32. doi: 10.1177/000312240707200402.

Chan, Tak Wing, and John H. Goldthorpe. 2007b. 'Class and Status: The Conceptual Distinction and Its Empirical Relevance'. *American Sociological Review* 72(4):512–32. doi: 10.1177/000312240707200402.

Chan, Tak Wing, and John H. Goldthorpe. 2007. 'Social Status and Newspaper Readership'. *American Journal of Sociology* 112(4):1095–1134. doi: 10.1086/508792.

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. doi: 10.1038/s41586-022-04996-4.

Choi, Hyunyoung, and Hal Varian. 2012. 'Predicting the Present with Google Trends'. *Economic Record* 88(s1):2–9. doi: 10.1111/j.1475-4932.2012.00809.x.

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. doi: 10.1038/s41598-021-87971-9.

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. doi: 10.1177/1461444813487959.

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.

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. doi: https://doi.org/10.1111/asap.12001.

DiMaggio, Paul, and Filiz Garip. 2012. ‘Network Effects and Social Inequality’. *Annual Review of Sociology* 38(1):93–118. doi: 10.1146/annurev.soc.012809.102545.

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* 255(5047):946–52. doi: 10.1126/science.1546291.

Duncan, Otis Dudley. 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. doi: 10.1126/science.1186605.

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. doi: 10.1080/0267257X.2014.989890.

Erikson, Robert, and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford [England] : New York: Clarendon Press ; 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. doi: 10.1016/j.worlddev.2018.03.007.

Filho, Renato Miranda, 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*. 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. doi: https://doi.org/10.1111/1468-4446.12295.

Flemmen, Magne Paalgard, 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. doi: https://doi.org/10.1111/1468-4446.12508.

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. doi: 10.1163/15691330-12341258.

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):1397–1436. doi: 10.1086/657976.

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. doi: 10.1177/0003122416632787.

Goldthorpe, John H., Catriona Llewellyn, and Clive Payne. 1987. *Social Mobility and Class Structure in Modern Britain*. 2nd ed. Oxford [Oxfordshire] : New York: Clarendon Press ; 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*. CRC Press.

Guess, Andrew M., Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. ‘The Consequences of Online Partisan Media’. *Proceedings of the National Academy of Sciences* 118(14). doi: 10.1073/pnas.2013464118.

Hargittai, Eszter, and Amanda Hinnant. 2008. ‘Digital Nequality: Differences in Young Adults’ Use of the Internet’. *Communication Research* 35(5):602–21. doi: 10.1177/0093650208321782.

Hauser, Robert M., and John Robert Warren. 1997a. ‘Socioeconomic Indexes for Occupations: A Review, Update, and Critique’. *Sociological Methodology* 27:177–298.

Hauser, Robert M., and John Robert Warren. 1997b. ‘Socioeconomic Indexes for Occupations: A Review, Update, and Critique’. *Sociological Methodology* 27(1):177–298. doi: 10.1111/1467-9531.271028.

Hinds, Joanne, and Adam N. Joinson. 2018. 'What Demographic Attributes Do Our Digital Footprints Reveal? A Systematic Review'. *PLOS ONE* 13(11):e0207112. doi: 10.1371/journal.pone.0207112.

Hodas, Nathan Oken, 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* 249–57. doi: 10.1109/SocialCom-PASSAT.2012.129.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. 'Prediction and Explanation in Social Systems'. *Science* 355(6324):486–88. doi: 10.1126/science.aal3856.

Holt, Douglas B. 1998. 'Does Cultural Capital Structure American Consumption?' *Journal of Consumer Research* 25(1):1–25. doi: 10.1086/209523.

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. doi: 10.1126/science.aaf7894.

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. doi: 10.1080/15230406.2018.1434834.

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. doi: 10.1109/ACCESS.2020.2983656.

Katz-Gerro, Tally. 1999. 'Cultural Consumption and Social Stratification: Leisure Activities, Musical Tastes, and Social Location'. *Sociological Perspectives* 42(4):627–46. doi: 10.2307/1389577.

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. doi: 10.1073/pnas.1218772110.

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. doi: 10.1177/0003122419877135.

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. doi: 10.1146/annurev.publhealth.18.1.341.

Kwon, Eun Sook, Eunice Kim, Yongjun Sung, and Chan Yun Yoo. 2014. 'Brand Followers'. *International Journal of Advertising* 33(4):657–80. doi: 10.2501/IJA-33-4-657-680.

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. doi: 10.1038/s41586-021-03660-7.

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. doi: 10.1126/science.1167742.

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). doi: 10.1098/rsif.2016.0598.

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):9. doi: 10.1007/s13278-018-0486-1.

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. 'Social Media Fingerprints of Unemployment'. *PLoS ONE* 10(5). doi: 10.1371/journal.pone.0128692.

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. doi: 10.1038/ncomms15227.

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. doi: 10.2307/2095397.

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. doi: 10.1177/0049124115605339.

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. doi: 10.1177/1461444818822816.

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. doi: 10.1016/j.jpubeco.2003.10.005.

Moseley, Nathaniel, Cecilia Ovesdotter 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. doi: 10.2307/2091776.

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. doi: 10.18637/jss.v020.i03.

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. doi: 10.1016/j.socnet.2017.06.002.

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. doi: 10.5465/AMBPP.2018.18881abstract.

Peterson, Richard A. 1992. 'Understanding Audience Segmentation: From Elite and Mass to Omnivore and Univore'. *Poetics* 21(4):243–58. doi: 10.1016/0304-422X(92)90008-Q.

Peterson, Richard A., and Roger M. Kern. 1996. 'Changing Highbrow Taste: From Snob to Omnivore'. *American Sociological Review* 61(5):900–907. doi: 10.2307/2096460.

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. doi: 10.1371/journal.pone.0138717.

Prieur, Annick, and Mike Savage. 2013. 'Emerging Forms of Cultural Capital'. *European Societies* 15(2):246–67. doi: 10.1080/14616696.2012.748930.

Reeves, Aaron. 2019. 'How Class Identities Shape Highbrow Consumption: A Cross-National Analysis of 30 European Countries and Regions'. *Poetics* 76:101361. doi: 10.1016/j.poetic.2019.04.002.

Rodríguez-Hernández, Carlos Felipe, Eduardo Cascallar, and Eva Kyndt. 2020. 'Socio-Economic Status and Academic Performance in Higher Education: A Systematic

Review'. *Educational Research Review* 29:100305. doi: 10.1016/j.edurev.2019.100305.

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 Bohme Carnegie, Ryan James 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 Lee 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):8398–8403. doi: 10.1073/pnas.1915006117.

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. doi: 10.1177/0038038513481128.

Savage, Mike, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013b. 'A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment'. *Sociology* 47(2):219–50. doi: 10.1177/0038038513481128.

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. doi: 10.3102/00346543075003417.

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. doi: 10.1371/journal.pone.0115545.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. ‘Limits of Predictability in Human Mobility’. *Science* 327(5968):1018–21. doi: 10.1126/science.1177170.

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. doi: 10.15195/v7.a23.

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. doi: 10.2139/ssrn.3144139.

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).

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. doi: 10.1089/cyber.2014.0609.

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)*.

de Vries, Robert, and Aaron Reeves. 2021. ‘What Does It Mean to Be a Cultural Omnivore? Conflicting Visions of Omnivorousness in Empirical Research’. *Sociological Research Online* 13607804211006108. doi: 10.1177/13607804211006109.

Wagner, Claudia, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. ‘Measuring Algorithmically Infused Societies’. *Nature* 595(7866):197–204. doi: 10.1038/s41586-021-03666-1.

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. doi: 10.1073/pnas.1418680112.

Yu, Jingyuan, and Juan Muñoz-Justicia. 2020. ‘A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science’. *Future Internet* 12(5):91. doi: 10.3390/fi12050091.

Chapter 4 Omnivorous, Inconspicuous, and Niche: High Socioeconomic Status is Associated with Diverse Consumption

Abstract

Combining insights from sociology, social psychology, and consumer research, we integrate the cultural omnivorousness and inconspicuous consumption theories and argue that high socioeconomic status is associated with more diverse consumption practices for all forms of consumption. Consumption practices are determined by a combination of economic, social, and cultural forces. This bundling dictates that lower economic constraints leave more room to diversify consumption along cultural and social aspects in the form of omnivorous or lifestyle-based niche consumption. We provide empirical evidence for the diversity hypothesis by analysing millions of mobile-tracked visits from thousands of Census Block Groups to thousands of stores in New York State. The results show that high income is significantly associated with diverse consumption across brands and price levels. The associations between diversity and income persist in different categories of consumption but are less prominent in the densely populated and demographically diverse New York City. The associations replicate for education as an alternative measure of socioeconomic status and for the state of Texas. We further illustrate that the associations cannot be explained by simple geographic constraints, including geographic mobility of the residents and local availability of the stores, so deeper social and cultural factors must be at play.

Introduction

An important aspect of socioeconomic inequality is the difference in daily consumption practices by socioeconomic status (SES). This difference is not only a manifestation of inequality, but also a trigger for further inequality in other life outcomes. For example, constrained by availability and price, people in low SES tend to go to low-price supermarkets and consume unhealthy food and beverages, which could contribute to later health problems (Baumann, Szabo, and Johnston 2019; Pechey and Monsivais 2015; Zagorsky and Smith 2017). Also, in more unequal societies, people tend to pay more attention to and spend more money on status goods (Heffetz 2011; Walasek, Bhatia, and Brown 2018; Walasek and Brown 2015), which may make them work more (Bowles and Park 2005), save less (Wisman 2009), accrue more debt (Christen and Morgan 2005), or even declare bankruptcy (Perugini, Hölscher, and Collie 2016).

Consumption is an economic, social, and cultural phenomenon that attracts research from multiple disciplines. From an economic perspective, studying consumption inequality is an important complement to studying income inequality, as the basic utility function of individuals typically includes consumption and leisure, and not necessarily income (Attanasio and Pistaferri 2016). Existing evidence suggests that short-term fluctuations in income inequality do not automatically transfer to consumption inequality due to the smoothing effect of savings, wealth, and borrowing, but persistent income inequality would be followed by consumption inequality. Consumption inequality itself leads to inequality in welfare and wellbeing and stagnant intergenerational mobility (Attanasio and Pistaferri 2016; Krueger and Perri 2006).

From the angle of social psychology, consumption practices are important because they express personal and social identities (Belk 1988; Reimer and Leslie 2004; Woodward 2003). Due to social comparison and status seeking, people often engage in consumption practices with the sole purpose to demonstrate how successful and rich they are. Veblen's ([1899] 2017) notion of *conspicuous consumption* captures the idea that some consumption is not economically practical but strategically employed to impress or show off to others. Research from this vein has shown that higher inequality exacerbates status competition and may thus lead to more interest in and consumption of status goods (Sahin and Nasir 2022; Walasek et al. 2018; Walasek and Brown 2015).

Complementing the economic and sociopsychological perspectives, sociologists bring in the cultural aspect of consumption. Bourdieu's (1984) theory of habitus posits that SES shapes taste, which in turn affects cultural consumption. People's consumption practices, particularly when they concern music, theatre, and other cultural events, encode their socioeconomic background and serve to distinguish their social position. In other words, cultural consumption both reflects and reproduces socioeconomic inequality.

This paper integrates insights from sociology, social psychology, and consumer research and presents new large-scale empirical evidence for socioeconomic inequality in daily consumption practices. We argue that since consumption practices are influenced by social and cultural factors, a reduction or removal of economic constraints fosters more differentiation along these aspects. We draw on two recent notions – cultural omnivorousness (Peterson 1992; Peterson and Kern 1996) and inconspicuous consumption (Berger and Ward 2010; Eckhardt, Belk, and Wilson 2015) – to argue that higher SES individuals engage in more omnivorous as well as more niche material consumption. As a result, higher SES is associated with more diverse consumption practices. We analyse mobile tracking data of US residents' visits to various stores to present evidence for this hypothesis. Our hypothesis extends the omnivorousness argument to all forms of consumption and integrates omnivorousness and niche consumption. Our findings illustrate and quantify socioeconomic divisions in daily consumption practices, bearing further evidence for the pervasiveness and inevitability of inequality in daily life.

The social and cultural process of consumption

Veblen's ([1899] 2017) theory of conspicuous consumption, introduced in the book *Theory of the Leisure Class*, is one of the earliest and most influential theories that describe the social process of consumption. According to Veblen, the economic and technological development of a society contributes to the evolution of a leisure class, whose members enjoy the surplus produced by the working class and do not need to work as much (Trigg 2001). The effective operation of surplus leads to the acquisition and accumulation of private property and thus the accumulation of property becomes an indicator of one's competence, which grants status and honour in the social hierarchy. As the association between property and status becomes widely accepted, the display of wealth evolves into a useful way to establish status. Veblen ([1899] 2017) suggests that the key to displaying wealth is showing the capacity to afford waste. He uses the notion of *conspicuous leisure* to describe the practice of displaying wealth through engaging in extensive leisure activities that basically waste effort and time, and the concept of *conspicuous consumption* to represent the practice of displaying wealth through purchasing and using goods that often cost more than their practical value. Moreover, Veblen argues that as mobility in society increases, it becomes harder to keep informed about others' leisure

activities, rendering conspicuous consumption a more effective display of wealth than conspicuous leisure.

Veblen regards conspicuous consumption as the most important determinant of consumer behaviour. He believes that as people in high social status display their wealth and status through conspicuous consumption, they also set up an ideal for those in the lower levels of the social hierarchy. Because people aspire to obtain status by consumption, people at each level of the social hierarchy emulate the consumption of those at a higher level; this is referred to by Veblen as *pecuniary emulation*. This process may never end. As those in lower status catch up, those in higher status must find out new goods of conspicuous consumption to distinguish themselves, creating new rounds of pecuniary emulation for the lower-status individuals. The concept of conspicuous consumption has proven useful in many cases. For example, it has been used to explain the choice of cosmetic brands (Chao and Schor 1998), purchase of niche products (Schaefers 2014), and racial differences in the proportion of expenditure spent on visible goods (Charles, Hurst, and Roussanov 2009). Conspicuous consumption also can be linked to the reproduction of inequality. Although conspicuous consumption has some positive effects such as the perception of elevated status and short-term happiness, it compromises the budget for basic needs and long-term progress and could lead to longer work hours, less savings, and more debt (Bowles and Park 2005; Christen and Morgan 2005; Kumar et al. 2021; Srivastava, Mukherjee, and Jebarajakirthy 2020; Wisman 2009).

While the research around conspicuous consumption mainly focuses on the social processes of economic consumption, sociologists bring in the cultural aspects of consumption. The most influential theory on this topic is Bourdieu's (1984) theory of cultural capital and taste, as part of a broader theoretical framework about socioeconomic inequality and its reproduction. Bourdieu (1984, 1986) views individuals' 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. The different forms of capital are correlated and transferable, but they still exist in different dimensions. According to Bourdieu, consumption is determined by not just economic capital but a combination of the different forms of capital. Bourdieu (1984) shows that in the 1970s, in France, not only cultural consumption such as literature and art, but also everyday consumption such as clothing and eating could be grouped by taste and correlated with SES.

Specifically, Bourdieu emphasizes how cultural capital shapes taste, which then affects economic and cultural consumption. Cultural capital introduces the cultural aspect to socioeconomic inequality and connects education, the cultural industries, and stratification (Warde 2015). As cultural capital is obtained by upbringing and formal educational training, taste is more subtle and requires more effort to acquire than economic capital. Upbringing shapes cultural capital, which in turn affects educational attainment, and thus unequal cultural capital gets reproduced effectively (Bourdieu and Passeron 2000; DiMaggio 1982). Further, status distinction is achieved through taste not just by confirming one's own status group's taste as superior, but also by rejecting other groups' taste as inferior (Trigg 2001); this negativity aspect increases the exclusivity of taste. Therefore, consumption involves both the confirmation of one's superior taste and the rejection of others' inferior taste.

The most influential alternative to Bourdieu's theory of cultural capital is the concept of cultural omnivorousness, which focuses on cultural consumption. Bourdieu seems to assume a linear relationship between cultural capital and taste, where people with high cultural capital

enjoy highbrow (elite) arts such as classical music and people with low cultural capital like lowbrow (mass) arts such as country music. Using survey data on people's preferences for music genres, Peterson (1992) shows that while such highbrow-lowbrow relationship is true when ranking music genres based on the SES of their audience, closer inspection shows a different story. After decomposing the audience's SES for each genre, he reveals that people in higher SES show large interest in most of the genres, whereas people in lower SES show interest in only few genres (Peterson 1992; Peterson and Kern 1996). People in higher SES may have the resources and exposure to consume a large variety of cultural products, whereas people in lower SES may have limited resources and exposure to only a few types of cultural products. Therefore, Peterson (1992) proposes that the distinction of cultural consumption between people in higher and lower SES is along the axis of omnivore–univore instead of highbrow–lowbrow.

Although the pattern of cultural omnivorousness has been identified in many cultural contexts and the concept has gradually become dominant in cultural sociology, the debate remains active (Chan 2019; de Vries and Reeves 2021; Warde 2015). There is still empirical evidence for highbrow snobbish taste among elites instead of omnivorousness (Atkinson 2011; Veenstra 2015). There are also vibrant debates about the extent to which cultural omnivorousness conflicts with Bourdieu's theory of cultural capital and what it means for the relation between culture, class, and power (Chan 2019; Goldberg 2011; de Vries and Reeves 2021). Notably, one may view omnivorousness itself as a new form of cultural capital, which makes the concept of cultural omnivorousness a complement rather than an alternative to Bourdieu's theory (Coulangeon 2017; Warde, Wright, and Gayo-Cal 2008).

The development of cultural capital as a concept also challenges and complements the notion of conspicuous consumption. Recent marketing research discovered that people with higher SES, especially those with higher cultural capital, tend to prefer inconspicuous consumption to conspicuous consumption (Berger and Ward 2010; Eckhardt and Bardhi 2020; Eckhardt et al. 2015). Inconspicuous consumption is marked by the subtlety of status signals (Berger and Ward 2010; Eckhardt et al. 2015) and is also related to the appreciation of experience and authenticity (Eckhardt and Bardhi 2020). Bourdieu's theory of cultural capital and taste provides a reasonable explanation for the shifting status signals of consumption. As conspicuous consumption becomes more affordable, people with higher SES need to reject the popular taste and develop new taste, which manifests as the appreciation of inconspicuous, experiential, or authentic goods and experiences.

Diverse consumption by SES

Although researchers of inconspicuous consumption draw inspiration from Bourdieu's cultural capital, they neglect the possible links with cultural omnivorousness. Meanwhile, researchers of omnivorousness mainly focus on the consumption of cultural products. In this paper, we aim to bridge the gap between the two concepts and argue that they stem from the same socio-psychological process and result in the same social pattern.

To summarize, existing research on consumption suggests that it is as much an economic as social and cultural phenomenon (Arnould and Thompson 2005). We argue, however, that social and cultural factors dominate consumption mainly when economic constraints fall. For high-SES individuals, consumption practices are less driven by product prices and more influenced by social and cultural processes of distinction. Notably, there are two ways in which distinction may occur.

On the one hand, distinction could be in reference to low SES. For people in high SES, high economic capital removes constraints of resources, high cultural capital provides the ability to be open-minded and appreciate diversity, and high social capital offers wider exposure to various consumption practices. These conditions have been used to explain cultural omnivorousness, but they hold true for material as much as for cultural consumption. Hence, to display broad-mindedness, progressiveness, and non-materialism, some high-SES individuals may engage in omnivorous consumption practices, consuming across the spectrum of brands and prices.

On the other hand, distinction may be confined within the high-SES stratum. People of high SES have the capacity to use consumption to demonstrate dedication to a lifestyle or ideology such as luxury and exclusivity, but also environmental sustainability, healthy and natural living, New Age beliefs, anti-globalization, even anti-consumerism in the form of inconspicuous, experiential, or authenticity consumption. In contrast, low-SES individuals who pinch pennies to provide bare essentials such as food, housing, and fuel do not have this privilege. Thus, to affirm their identity and lifestyle, some high-SES individuals may consume within market niches, bundling brands and services with intention and purpose.

While omnivorousness implies that high-SES individuals engage in diverse consumption practices, lifestyle consumption implies that high-SES individuals consume within narrow niches but that there is a diversity of niches. Whether taken individually or together, both processes imply that, in the aggregate, high SES is associated with more diverse consumption practices. This is the hypothesis we test here.

Prior research already provides partial evidence in support of omnivorous and niche consumption among high-SES individuals. The qualitative research on inconspicuous consumption among elites we already referred to corroborates the niche consumption argument (Berger and Ward 2010; Eckhardt and Bardhi 2020; Eckhardt et al. 2015). From computer science, one study that linked mobile phone with banking transaction data from Mexico found strong associations between purchase patterns and SES groups (Leo et al. 2018). The researchers found that high SES is correlated with more diverse purchases across product and service categories (Leo et al. 2018) and merchants (Dong et al. 2020), offering large-scale evidence for omnivorousness in material consumption.

Here, we extend this work in three ways. First, we make a theoretical contribution whereby we generalize about the association between SES and diversity in consumption to encompass omnivorousness, niche, and inconspicuous consumption. On the one hand, we extend the concept of omnivorousness from cultural to any consumption, and on the other, we bring attention to the problem of inequality to studies of lifestyle, niche, and inconspicuous consumption. By embedding our arguments and empirical observations into several established areas of research, we hope to stimulate further theoretical elaboration and empirical investigation in the social sciences. Second, we replicate the previous empirical findings on product categories from Mexico for brands and in the US, confirming the universal nature and broad reach of the problem we study. Third, we provide evidence with data that are easily available for researchers at a reasonable fee, in contrast to the restricted, sensitive, and private individual banking and communication data studied by Leo and colleagues (2018). Our data are at the aggregate and not the individual level, but they are easily accessible, allowing the wider scientific community to replicate and extend the findings presented here.

Methods

Data

The data for this paper come from three sources: SafeGraph, the US Census Bureau, and Yelp. SafeGraph is a company that curates geospatial data linked with mobile tracking data for a large panel of US smartphones (SafeGraph 2023). In the main datasets from SafeGraph, each observation is a place that is a point of interest (POI), namely, a specific physical location that someone may find interesting (restaurants, retail stores, grocery stores, etc.). For each POI, SafeGraph provides a range of information including the store's name, street address, counts of visits, the home census block groups (CBGs) of the visitors, the store's North American Industry Classification System (NAICS) category, the brand of the store, and other brands that the visitors visited on the same day/week/month, etc. SafeGraph provides data on visits to the stores of more than 7,000 distinct brands, covering all the major brands in the US. Around 80% of POIs do not have an associated brand, as they are unique commercial locations (local restaurants, museums, etc.). In this paper, we only work with POIs that are associated with brands and aggregate data of POIs to the brand level.

The NAICS codes are hierarchical six-digit codes used to represent the industries, where the first two digits represent the most general categories, and the full six digits represent the most specific categories. We selected 28 four-digit categories that are related to daily consumption and these contain 78 six-digit categories. We also filtered by location and time. We present results based on data from New York State in October 2019. We chose October 2019 because it is the most recent month before the COVID-19 pandemic that is not affected by holiday shopping seasons. We selected to focus on one state rather than the entire country since many brands are regional. New York State is a reasonable choice because it is populous, ranking fourth among US states with a population of nearly 20 million, and the most socioeconomically unequal, with income Gini coefficient of 0.51 (US Census Bureau 2022). This guarantees enough observations and variability in the data. We have no reason to expect that the phenomenon and mechanisms we investigate here differ qualitatively for other states but, for robustness, we replicated the findings with data from another populous state, Texas (see Appendix E).

After the filtering, we have 264,826 observations in total. Table A1 in Appendix A shows the four-digit category names, the NAICS codes, and the number of POIs for each category. Among the 264,826 observations, 35,588 (about 13 percent) have brands associated with them. After dropping observations with limited information, we have 24,188 POIs that belong to 1175 brands. Among the 28 categories, four of them ("Gambling Industries", "Museums, Historical Sites, and Similar Institutions", "Performing Arts Companies", "Special Food Services") do not have any brand, so we end up with 24 categories.

For 23,536 of the 24,188 POIs, each observation has a table of the visitors' home CBGs and the number of visitors from each CBG. The CBG is the smallest geographic unit for which the US Census Bureau publishes sample data; it normally has population of 600 to 3000 people. The visitors to the 23,536 POIs are from 54,307 CBGs. SafeGraph provides census data from the US Census Bureau's American Community Survey 5-year Estimates (SafeGraph 2022; US Census Bureau 2022). We downloaded the 2019 census data and linked the CBGs with the median household income of each CBG. Not all CBGs have available median household income estimates; we were able to link 52,509 CBGs in our sample. Aggregating the data by brands, we get a bipartite network of 52,509 CBGs and 1,150 brands, where the weights are

the number of visits from each CBG to each brand. As SafeGraph does not have perfect data and we are also limiting the visits in terms of location and time, we left out data with insufficient number of observations. From the 1,150 brands, we dropped the brands that have less than 100 incoming visitors in our sample in the month, resulting in 924 brands. From the 52,509 CBGs, we dropped the CBGs that have less than 100 outgoing visitors in our sample, resulting in 13,653 CBGs. Compared with all 52,509 CBGs, the selected 13,653 CBGs slightly oversample lower income CBGs, but this factor is not anticipated to exert any discernible influence on the results of our study (Figure A1).

We used Yelp to get an indicator for the price level of the brands. Yelp is a social media platform that publishes crowd-sourced reviews of businesses. It is primarily used for restaurant reviews, but also includes reviews of other businesses such as clothing stores, department stores, and grocery stores. For many stores, Yelp provides a reference price level in the form of dollar signs ($) that indicates the average cost per person. From one to four dollar signs, the price levels are under $10, $11-$30, $30-$60, and above $60. Yelp officially provides an open dataset that covers businesses in 11 metropolitan areas (Montreal, Calgary, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland). With text matching, we were able to match 371 brands in our sample with the Yelp Open Dataset. For the few brands that have stores with different price levels, we use the mode price level. For the rest of brands, we manually searched the brands on Yelp, setting the location in New York City and other cities in the Net York State. Combining the data from the Yelp Open Dataset and manual searching, we were able to find the Yelp price levels for 783 brands in our sample.

Statistical Analyses

We analyse the data from two angles. We first use the brands as unit of analysis. For each of the 924 brands, we use the median household income of the visitors' home CBG to obtain a distribution of the median household income of brand visitors. We then use the median value from the distribution to represent the typical household income of the brands' visitors, which can be considered an indicator of the brands' SES. We compare the typical income of brands' visitors and the price level of the brands to examine the extent to which brands' visits are determined by the price and we analyse the outliers to identify different consumption patterns.

Second, we use the CBGs as unit of analysis. To test our main hypothesis, we explore the correlations between the median household income of the CBGs and three measures of the diversity of consumption. The first measure uses the standard deviation of the typical income of the brands' visitor to test whether people living in CBGs with higher median household income tend to visit more diverse brands in terms of SES. This measure is somewhat tautological as the brand SES is constructed by the visitors' CBGs, so the interpretation should be cautious. The second and third measures use a function of Shannon entropy, which is a common diversity measure in various contexts. The second measure is for brands and the third measure is for the brands' price levels.

We use the normalized Shannon entropy. For a CBG *i*, the normalized Shannon entropy is:

$$D(i) = \frac{-\sum_{j=1}^{k} p_{ij} \log\left(p_{ij}\right)}{\log\left(k\right)}$$

where k is the number of brands or price levels in our sample and p_{ij} is the proportion of visits from CBG *i* to brand/price level *j* out of the total number of visits of *i*. For the entropy by

brands, k = 924 and for the entropy by price level, k = 4. For robustness, we replicate the analyses with three additional measures of diversity (Table A2). To provide indirect evidence for the supposed mechanisms, we also disaggregate the analyses by industry and for New York City versus the rest.

To reject alternative explanations for the observed affects, we use regression analyses to predict CBG diversity in consumption with median household income, while controlling for CBG residents' mobility and local brand availability. We measure mobility with the median distance travelled from the centre of the CBGs to the stores. We measure local availability with the same measures we used for consumption diversity, but with the brands available within the CBG instead of the brands visited. We run the regression model for all data and for each industry separately, including fixed effects for CBGs located within New York City versus outside. To address the potential ecological fallacy due to the aggregate analysis, we also control for available demographic variables for the CBGs, including median age, proportion of male residents, proportion of white residents, and proportion of residents whose highest degree earned is a bachelor's degree or higher.

Network Embedding

To test for niche consumption patterns, we use data on brand co-visits. We normalize the data such that we obtain a 924x924 matrix where each *i,j* entry is the proportion of visitors from brand *i* who also visited brand *j* within that month.

To analyse the data, we first use the graph embedding method node2vec (Goyal and Ferrara 2018; Grover and Leskovec 2016). This is a semi-supervised method that learns a mapping of nodes (the brands, in this case) to a lower-dimensional vector space. Distances in this new embedding space reflect neighbourhood similarity in the original network. The method uses a biased random-walk procedure to explore each node's neighbourhood. The procedure relies on two parameters, p and q, which control the extent to which the neighbourhood is defined locally, in the sense of a proximate community, or globally, in the sense of the network concept of structural equivalence. We set low p = 1 and larger q = 2 in order to map nodes by proximity in the network – the more relevant concept for the definition of niche consumption.

After compressing the matrix to 128 dimensions, we visualize the results with t-SNE (t-distributed stochastic neighbour embedding) in two dimensions. This method depicts nodes that are similar in the 128 dimensions as points that are closer together on the plot than nodes that are dissimilar. Additionally, we identify clusters by applying k-means clustering on the 128-dimension embedding. This method groups the nodes into k clusters by assigning each node to the cluster whose current centroid is closest. We attempted to use the elbow method to determine the number of clusters k but since no clear elbow transpired, we chose the clustering that mapped best onto the t-SNE visualization.

We note that we explored various parameters, data processing, and methods. Specifically, we tested different combination of parameters for node2vec: number of dimensions = (32, 128, 256); p = (0.25, 0.5, 1), and q = (1, 2, 64). We also tested different python implementations of node2vec, as well as additional algorithms: the Walktrap community detection algorithm (Pons and Latapy 2006) and structural deep network embedding, or SDNE (Wang, Cui, and Zhu 2016). Finally, we also ran these analyses on differently trimmed network, with minimum threshold of ties for 1, 5, and 10 co-visits. We do not report the results from these additional analyses here as they are not notably different.

Replicability

To obtain a more comprehensive understanding of consumption inequality by SES, we replicate the analyses with education as an alternative proxy for SES. From the census data available, we measure education as the proportion of residents who are 25 years and over in the CBGs and whose highest educational attainment is a bachelor's degree or higher. We also try average years of education, where we calculate the weighted sum of the proportion of residents with different highest educational attainment multiplied by the years of education required. The average years of education measure is highly correlated with the proportion of bachelor's degree or higher (0.888, $p < 0.001$) and the findings are the same, so we only report the results by the proportion of bachelor's degree or higher (Appendix D).

Finally, to test the external validity of our findings, we replicate the analyses with data from another populous state, Texas, taken over the same time period (reported in Appendix E).

Results

Descriptive results

As expected, SES is associated with different consumption preferences for consumer brands. For instance, using a simple LASSO regression model (reported in Appendix B), we can predict the median household income of the CBGs with the proportion of outgoing visitors of the CBGs to the 924 brands, obtaining an out-of-sample correlation between predicted and actual median CBG income of 0.748 ($p < 0.001$). This association does not necessarily correspond to economic constraints driven by the product prices. Figure 1 shows the distribution of brands' SES (measured by the median income of brand visitors) for different Yelp price levels, with some typical brands labelled. As the plot shows, in general, pricier brands have higher median income of visitors, but there are large variations between brands in the same price level. Some brands perform as expected. For example, discount stores such as Save-A-Lot and Price Rite have both low price levels and median income of visitors, while expensive supermarkets such as Whole Food and luxury fashion brands such as Valentino have high price levels and high median income of visitors. Some brands show surprising patterns. For example, cheap supermarket Lidl has relatively high median income of visitors whereas luxury fashion brand Gucci has relatively low median income of visitors.

Figure 2 shows the distribution of brand visitors' income for some typical brands identified from Figure 1 and, for reference, the distribution of median household income for the CBGs in our sample. As cheap grocery and department stores, Save-A-Lot and Sears have the expected distributions. But the cheap grocery and clothing stores Lidl and Gap attract large proportion of middle to high income visitors, showing possible patterns of inconspicuous or omnivorous consumption. Whole Foods and Valentino have large proportion of middle to high income visitors, which is expected, but they still attract low-income visitors, which may indicate conspicuous consumption. Gucci and the expensive cosmetics brand MAC Cosmetics have surprisingly large proportion of low to middle income visitors, showing even stronger possibility of conspicuous consumption. Both luxury fashion brands, Valentino and Gucci exhibit a dramatic difference in visitors' income distribution.

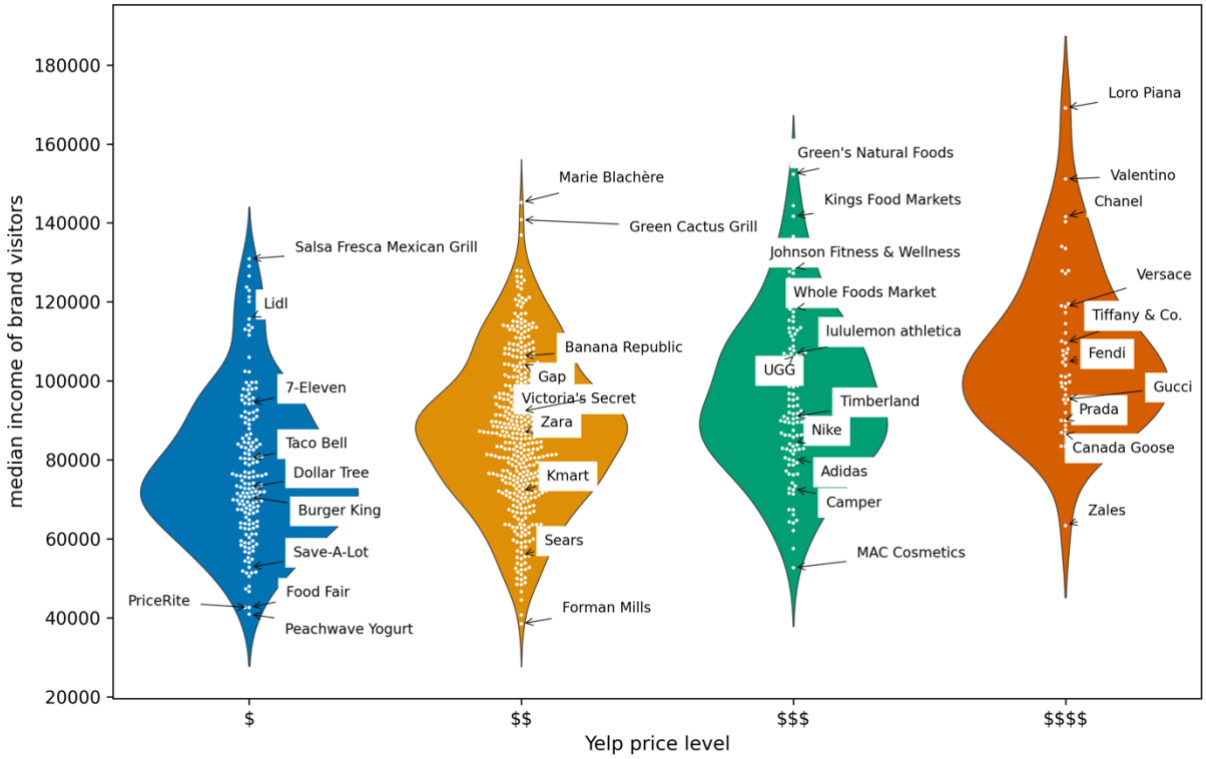

Figure 1. Relation between brands' Yelp price level and median income of brand visitors.
Note: Two extreme outliers, Learning Express Toys (median income 203,438; Yelp price level $$) and Balduccis (median income 250,001; Yelp price level $$$), are excluded for better visualisation.

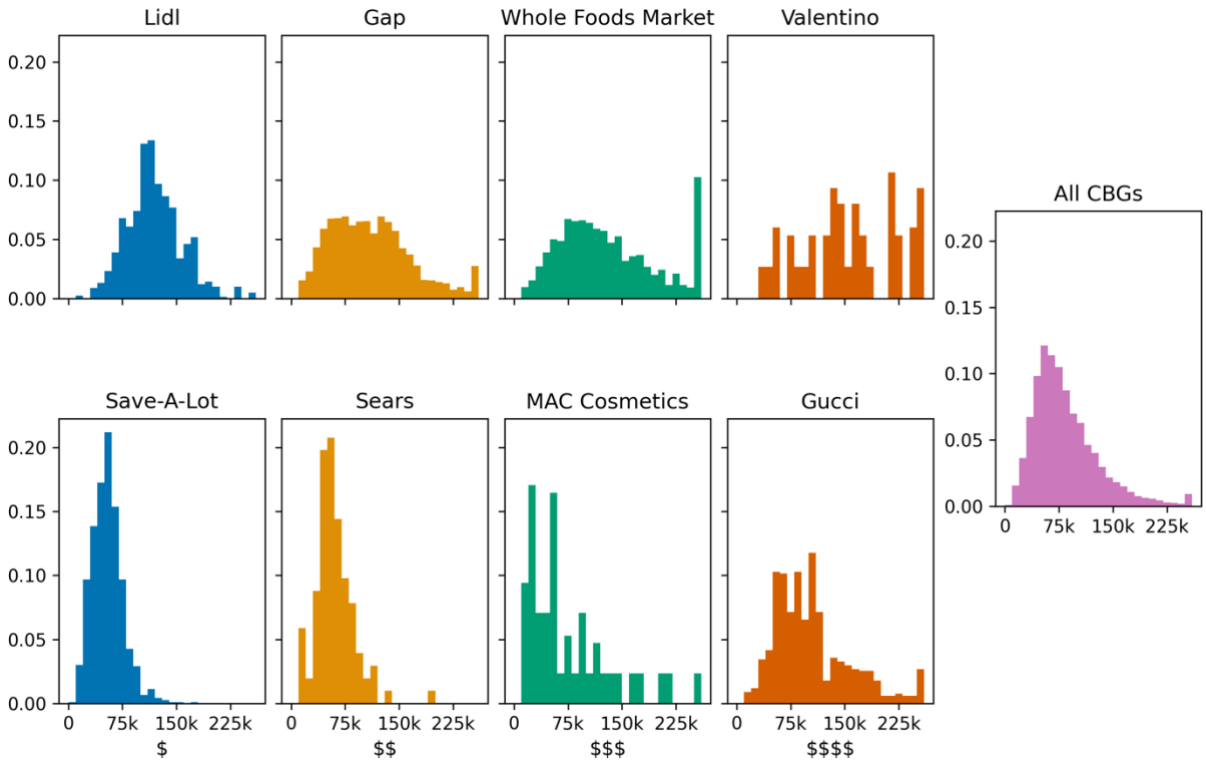

Figure 2. Distribution of brand visitors' income for some typical brands and all CBGs in the sample.

Diversity in consumption

Next, we find support for the consumption diversity hypothesis. All measures of diversity have significant and positive correlation with CBGs' median household income. Figure 3 shows the correlation between CBGs' median household income and three measures of diversity: a) the Shannon entropy by brand, b) the standard deviation of visited brands' SES, and c) the Shannon entropy by brands' price level. The correlation coefficients are 0.292, 0.471, and 0.358, respectively; all correlations are statistically significant at the 0.001 level. In brief, these results indicate that people residing in CBGs with higher median household income tend to visit more diverse brands, brands of more diverse SES, and brands of more diverse price levels.

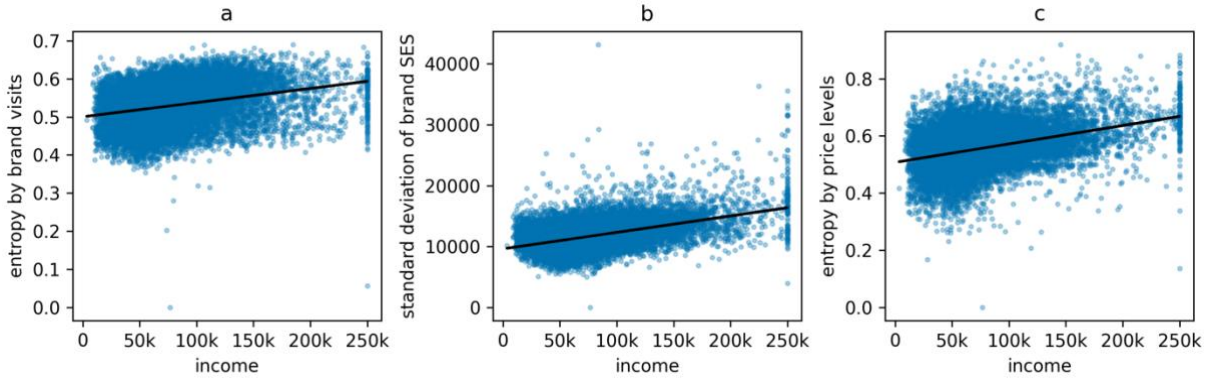

Figure 3. Correlation between CBGs' median household income and three measures of consumption diversity: a) Shannon entropy of brand visits; b) standard deviation of brand SES; c) Shannon entropy of brand price levels.

To further test the diversity hypothesis and obtain a more nuanced understanding of the observed patterns, we disaggregate the analysis by industry and investigate how the associations between SES and diversity vary for different industries. We use the three-digit NAICS codes to group similar industries together. Table 1 shows the association between CBGs' median household income and the diversity measures by industry. There are significant associations between income and the diversity measures in all the industries, confirming the robustness of the diversity phenomenon. It also appears that the association is stronger in industries that involve more cultural aspects (e.g., *Food Services and Drinking Places, Clothing and Clothing Accessories Stores, Amusement, Gambling, and Recreation Industries Sporting, Goods, Hobby, Musical Instrument, and Book Stores, Miscellaneous Store Retailers*) than those that concern necessity goods (e.g., *Food and Beverage Stores, Health and Personal Care Stores, General Merchandise Stores, Gasoline Stations*). The category *Motion Picture and Video Industries* is an exception as we only have nine cinema chain brands here; we expect that, in reality, the diversity comes more from independent cinemas. These differences do not appear to be driven by differences in brand and price variability between industries (last three columns in Table 1). Overall, the pattern corresponds well with the rationale of the diversity hypothesis: higher prominence of social and cultural factors should make omnivorousness and/or niche consumption more salient.

Table 1. The associations between CBGs' median household income and diversity in consumption by industry.

Industry	Associations with income			Industry characteristics		
	Entropy by brand	Std of brands' SES	Entropy by brands' price level	Number of brands	Std of brands' SES	Std of brands' price level
Amusement, Gambling, and Recreation Industries	0.329***	0.246***	0.153***	89	19314	1.035
Miscellaneous Store Retailers	0.271***	0.216***	0.083***	49	23878	0.598
Clothing and Clothing Accessories Stores	0.239***	0.098***	0.232***	203	19654	0.768
Food Services and Drinking Places	0.238***	0.444***	0.404***	297	18584	0.663
Sporting Goods, Hobby, Musical Instruments, and Book Stores	0.222***	0.254***	0.042***	53	25145	0.592
Motion Picture and Video Industries	0.111***	0.074***	–	9	21571	–
Health and Personal Care Stores	0.108***	0.022*	0.067***	35	20154	0.512
Gasoline Stations	0.103***	0.167***	0.036***	47	14221	0.494
Food and Beverage Stores	0.070***	0.064***	0.100***	95	30776	0.640
General Merchandise Stores	0.050***	0.130***	0.210***	42	21078	0.935

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests). Industries are ordered in descending order by entropy by brand. Yelp price levels are not available for brands in *Motion Picture and Video Industries. Personal and Laundry Services* and *Rental and Leasing Services* are excluded due to limited number of brands.

In order to explore how urban-rural lifestyle differences influence consumption diversity, next we disaggregate the analysis for New York City (NYC) and the rest of the state. We study the following cases: visits from CBGs in NYC to stores in New York state, visits from CBGs outside of NYC to stores in New York state, visits from CBGs in NYC only to stores in NYC, and visits from CBGs outside of NYC to stores outside of NYC. Table 2 shows the associations between CBGs' income and the diversity measures for those cases. The associations between income and diversity are much stronger for people who live outside NYC than people who live in the city. This suggests that NYC dwellers engage in equally diverse consumption regardless of their income. Some of this might be explained with the high density and diversity of consumption options in the city, the weaker constraints imposed by mobility, and possibly, the stronger social comparison and influence imposed by the denser population.

Table 2. The associations between CBGs' median household income and diversity in consumption by CBGs in and outside New York City (NYC).

Region		Entropy by brand	Standard deviation of brands' SES	Entropy by brands' price level
NYC CBGs	visiting all stores	-0.038**	0.130***	0.201***
	visiting only NYC stores	0.122***	0.022	0.040**
Non-NYC CBGs	visiting all stores	0.476***	0.603***	0.527***
	visiting only non-NYC stores	0.380***	0.599***	0.506***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).

Robustness to alternative explanations

Next, we examine the extent to which the main finding that high SES is associated with diverse consumption practices can be explained by simple geographic constraints. One such constraint is geographic mobility – high SES individuals may have higher geographic mobility and thus access to more consumer choices. Another constraint is local availability – high-SES CBGs may have more diverse consumer options due to urban planning or companies' strategic marketing. A third alternative explanation is that CBGs with high median income have higher variance of incomes and hence, our aggregate analyses simply capture the trivial fact that more diverse individuals consume more diversely. We find that median income is not significantly correlated with mobility and only weakly associated with local availability (0.075 for entropy by brand, 0.201 for brands' SES, 0.082 for entropy by price, $p < 0.001$ for all). Nevertheless, income is strongly correlated with income variability (0.592, $p < 0.001$), measured by the margin of error from the census data.

We include the factors in a regression model to predict consumption diversity by median income, controlling for income variability, mean distance travelled, and local availability, including fixed effects for CBGs in NYC. To mitigate the possibility of ecological fallacy, we also control for demographic characteristics of the CBGs including median age, proportion of male residents, proportion of white residents, and proportion of residents whose highest degree earned is a bachelor's degree or higher. The variables are standardised to compare the relative importance of the predictors. Table 3 shows the results from the regression models using different diversity measures. We find that median income is significantly associated with consumption diversity even after controlling for mobility, local availability, income variability, and the demographic variables, and it plays the most significant role among the predictors.

We repeat the regression analyses separately by industry (apart from *Motion Picture and Video Industries* due to limited number of observations with price and SES data). Regression results for each industry are available in Appendix C. The results are consistent for most industries. Income has a significant positive association with and is the strongest predictor of consumption diversity across all industries when we use entropy by brand as the measure. The same holds true using the other two measures with a few reasonable exceptions when the number of observations is limited or some cases where income and local availability have similar level of association with diversity. Overall, these results suggest that the association between income and consumption diversity is not likely to be mediated or confounded by mobility and local availability. The results provide further support for the assumptions behind the diversity hypothesis as they show that there are deeper social and cultural factors affecting consumption practices beyond simple geographic constraints.

Table 3. Results from regression analyses that predict CBGs' diversity in consumption with median household income, controlling for income variability, estimated mobility, estimated local availability, and demographic variables.

	Entropy by brand	Standard deviation of brands' SES	Entropy by brands' price level
Income	0.586***	0.363***	0.230***
	(0.019)	(0.021)	(0.019)
Income variability	-0.169***	0.032*	0.011
	(0.014)	(0.015)	(0.014)
Mobility	-0.207***	-0.168***	-0.085***
	(0.013)	(0.015)	(0.013)
Local availability	0.160***	0.226***	0.089***
	(0.011)	(0.013)	(0.011)
In NYC	-0.111***	-0.502***	0.501***
	(0.031)	(0.035)	(0.030)
Median age	0.040***	0.027*	0.057***
	(0.012)	(0.013)	(0.012)
Proportion of male	-0.022*	-0.022	0.010
	(0.011)	(0.012)	(0.011)
Proportion of white	-0.191***	-0.339***	-0.125***
	(0.016)	(0.018)	(0.015)
Proportion of bachelor's degree or higher	-0.161***	0.149***	0.213***
	(0.017)	(0.019)	(0.016)
Intercept	0.053***	0.184***	-0.173***
	(0.016)	(0.017)	(0.015)
Observations	6088	3672	5739
R^2	0.247	0.401	0.315
Adjusted R^2	0.246	0.399	0.314
Residual Std. Error	0.868	0.745	0.820
F Statistic	221.435***	272.347***	292.654***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).

Niche consumption

Aggregate data does not allow us to test for omnivorousness. Nevertheless, prior research using individual-level mobility and banking transaction data offers supporting evidence with regards to product and service categories (Leo et al. 2018) and merchants (Dong et al. 2014). Although individual-level data are also preferable to test for niche consumption, we can still use the aggregate data to test for expected macro-level patterns. Namely, if niche consumption is more common for high-SES individuals, then we will observe more numerous, smaller, and more clearly defined consumption communities for those from high-SES CBGs; these communities will also vary greatly in average price level. In contrast, those from low-SES CBGs will be visiting a larger grouping of brands at low price level.

The results of the network embedding analysis are presented in Figure 4. Apart from a couple of exceptions (e.g., the Pret A Manger/Bloomingdales cluster), the t-SNE plot does not identify distinct consumption niches. Additionally, mapping the 10 clusters identified by the k-means algorithm by mean SES and price does not reveal the expected pattern of niche consumption by SES. Instead of fewer and larger clusters at low SES, we observe the opposite – large consumer clusters for middle and upper-middle SES and more numerous and smaller clusters for low SES. The high-SES clusters display more price-level dispersion but the patterns are not that clear. Although we don't observe the aggregate co-visit patterns we expect from niche consumption, the definite test for it should be done with individual-level data.

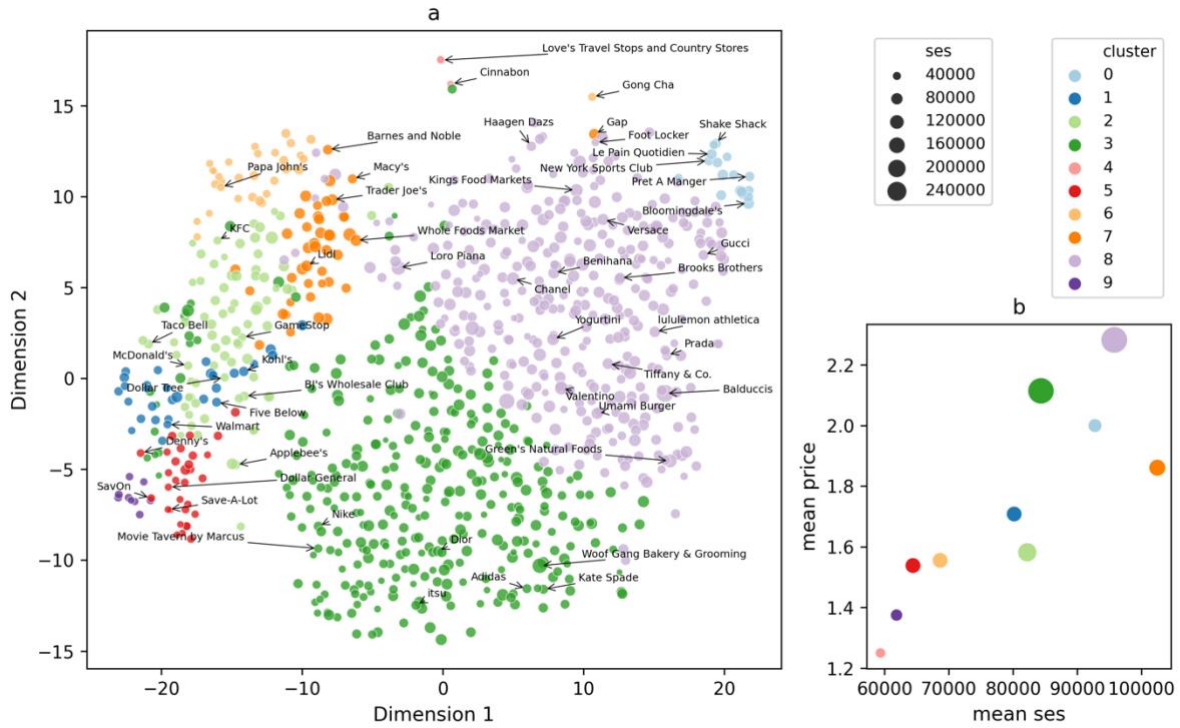

Figure 4. Niche consumption analysis: a) t-SNE visualization and k-means clustering of the brand co-visit network after node2vec embedding to 128 dimensions; b) mean SES and price for the brands in each cluster.
Note: In panel a, the marker size represents the brand's SES, as the legend indicates. In panel b, the marker size is proportional to the square root of the number of brands in each cluster.

Replicability

Overall, the results replicate whether we use education as a measure of SES or use data from Texas.

The proportion of people who have bachelor's degree or higher is significantly correlated with the three measures of diversity: the correlation coefficients are 0.133 for entropy by brand visits, 0.381 for the standard deviation of visited brands' SES, and 0.390 for entropy by brands' price level; all correlations are statistically significant at the 0.001 level (see Figure D4). As shown in Table 3, controlling for income, income variability, mobility, local availability and other demographic variables, education has a significantly positive association with diversity measured with the standard deviation of visited brands' SES and the entropy by brands' price level, but the association is negative when diversity is measured with the entropy by brand

visits. To address this inconsistency, we introduce an interaction effect between income and education in our regression analyses, the results of which are presented in Table D3. There is a significant negative interaction effect. Figure D5 illustrates the interaction effects by showing the regression lines of income (or education) and diversity given different values of education (or income). In most cases, the associations between income or education and diversity are positive. The interaction suggests that the correlation between income (or education) and diversity is weaker for people with high education (or income). When measuring diversity with the entropy by brand visits, the correlation between education and diversity becomes negative for high income CBGs. The findings are similar if we repeat the regression analyses separately by industry (see Tables D4, D5, D6).

All analyses conducted in Texas reveal similar patterns to those observed in New York State (Appendix E). In fact, the data in Texas show stronger support for our main hypothesis. In Texas, both income and education have significant and positive correlation with all three measures of diversity, both in general and separately by industries. Notably, unlike in New York State, where the correlations between education and diversity turn negative for a few industries, the correlations remain positive in Texas. These positive associations persist even after controlling for income variability, mobility, local availability, age, gender, and race. The interaction effects between income and education vary but largely support our main hypothesis. Additionally, descriptive patterns and brand co-visit analyses in Texas mirror those observed in New York State.

In summary, the replication analyses, utilizing education as an alternative measure of SES and employing data from Texas, consistently support our primary finding that high SES is associated with diverse consumption patterns. However, these replication analyses also reveal nuanced variations within the phenomenon. Notably, in New York State, weak negative associations between education and diversity are observed for a few industries related to necessity goods, whereas in Texas, the same industries exhibit positive associations. The interaction effects of income and education on diversity are not uniformly clear, with varying directions and significance levels across states, diversity measures, and industries. These nuances, while not undermining our hypothesis, underscore the complexity of the phenomenon. It is hard to dig into these nuances with the data available for this paper. Future research with more detailed individual level data is needed to address these complexities.

Discussion and Conclusion

Using large-scale data of mobile tracked visits, our study reveals inequality in daily consumption: high-SES individuals consume more diversely in terms of brands and price levels than low-SES individuals. It is not true that expensive goods are for the rich and cheap goods for the poor. Rather, cheap goods are for the poor and all goods are for the rich. In other words, inconspicuous consumption, niche consumption, and brand omnivorousness are options and privileges mainly for the rich. We find that the association between diverse consumption and SES is prevalent across different industries, although stronger in industries that involve leisure and cultural expression compared to those that concern necessity goods. We further establish that the association cannot be attributed entirely to simple geographic constraints, suggesting deeper social and cultural factors.

Our contribution is both theoretical and empirical. First, combining insights from sociology, social psychology, and consumer research, we integrate the separate and to a certain degree

opposing theories of cultural omnivorousness, conspicuous consumption, and inconspicuous consumption into one coherent theoretical argument. Essentially, we extend the omnivorousness argument from cultural sociology beyond cultural consumption to any consumption and we subsume omnivoursness, niche, and inconspicuous consumption under the phenomenon of diverse consumption. We argue that high-SES individuals are more omnivorous but also more niche-focused in both their cultural and material consumption because the lack of economic constraints make social and cultural aspects of expression and distinction more salient.

Second, combining data from multiple sources, we offer large-scale empirical evidence for the hypothesized links between inequality and consumption diversity. Our research used data from one US state and replicated the results for another one, which gives us confidence that the findings will qualitatively hold in other states and countries. Although our analyses are aggregate and cannot differentiate between omnivorous and niche consumption, prior research leads us to expect that both contribute to produce the observed pattern. Our findings quantify the extent to which low-income individuals are constrained in their everyday choices, and the extent to which the cornucopia and freedom of choice of market economies do not benefit all.

Nevertheless, we acknowledge several limitations to our research. One source of potential bias is the fact that we focus on brands with multiple stores and ignore a large number of unique shops and institutions. Yet, this may not be necessarily problematic. It is reasonable to assume that high-SES individuals are more likely to visit boutique shops and unique institutions such as museums and galleries, which means that we underestimate how diverse their consumption is. This means that the effects we find are even stronger in reality. Similarly, our data concern visits to brick-and-mortar shops but not online shopping. This omission may be consequential for our findings only if low-SES individuals are more likely to shop online than in person compared to high-SES individuals and/or shopping patterns differ systematically between online and in-store.

Further, we acknowledge that mobile coverage and representation, on which the mobile-tracking data depend, likely increase with higher SES. Thus, the consumption diversity we observe could be just due to the fact that we track a larger number of high SES individuals. In other words, it is not that the average high-SES individual consumes across a wider range but that we are capturing a larger number of high-SES individuals who might have just as narrow but non-overlapping consumer ranges. However, while this undermines the omnivorousness argument, it does not necessarily challenge the niche-consumption explanation.

Our data track visits to physical locations, but we do not know whether and to what extent visitors complete purchases there. It is possible that a significant proportion of the records reflect “window shopping” and not actual consumption. If we assume that this is more likely to be the case for low-SES individuals, this means that we are overestimating actual consumption for that group which, once again, gives us even more confidence in our findings. Certainly, it is also possible that it is high-SES individuals who engage in more leisure shopping or “retail therapy.” However, this tendency may be counteracted by the fact that wealthy individuals are also more likely to have shopping assistants. Unless they are live-in staff, shopping assistants will in fact boost the diversity of consumption for low-SES census block groups, and thus bias our results towards smaller, rather than larger effects. Although not available to us, SafeGraph offer data on money spent at POIs and hence, future research could replicate our analyses of visits for purchases.

Using Yelp for price comparisons entails further limitations. Price levels on Yelp are crowdsourced and hence, subjective, noisy, and potentially biased. Users' evaluations likely depend on expectations about the industry and geographical region. For instance, what can be considered expensive in Horseheads, NY may be cheap in New York City, and what is inexpensive for a new pair of jeans may be pricey for a meal. However, from our theoretical perspective centred on social distinction, relative comparisons are preferable to absolute prices and subjective evaluations more valuable than unfamiliar objective price indices. Still, high-SES areas may have more visitors and tourists contributing Yelp reviews, and their evaluations may be upwardly biased compared to residents' evaluations. Despite this, the Yelp price level estimates largely replicate the results obtained with alternative measures of consumption diversity, affirming the robustness of our findings.

Most importantly, we remind the reader that the study is conducted at the aggregate level and may not necessarily reflect individual choices and behaviour. Averaging behaviour inevitably hides much nuance and detail but may also introduce bias. We analyse census block groups and hence, we lump together 600-3000 individuals into a single unit. However, these individuals are not independent since some of them cohabit in households. Moreover, these individuals may be quite heterogeneous and the heterogeneity may be higher for census block groups with higher median income. Although we control for income variability in the regression models, the data prevent us from distinguishing the extent to which the average high-status individual is diverse in their consumption (omnivorousness) from the extent to which the high-status strata comprise diverse individuals with narrow consumption preferences (niche consumption). Research with detailed individual consumption data already provides some evidence for omnivorousness (Leo et al. 2018) but we require more quantitative research of inconspicuous and niche consumption by SES. One promising direction is to extend the concepts of variety and atypicality regarding cultural preferences (Goldberg, Hannan, and Kovács 2016) to consumption in order to quantify the relative prevalence of omnivorous and niche consumption. Another promising direction is to use qualitative or survey research to test the two supposed pathways of between- and within- social class distinction that may be influencing consumption patterns.

Overall, our research suggests that inequalities in material consumption parallel inequalities sociologists have already established regarding cultural taste: low-SES individuals are more constrained in their consumption practices than high-SES individuals. Importantly, however, the constraints are not necessarily economic but possibly cultural and social; this explains why low-SES individuals frequent expensive Gucci stores, for instance. This has an important implication. The cultural and social aspects of daily consumption could entrench economic advantages and disadvantages: the rich save by choosing to consume at the lower end, while the poor get in debt by being tempted to consume at the higher end. Consequently, to fight inequality, we require levelling policies and nudges at the cultural and social, not just economic, levels.

Acknowledgements

The authors are grateful to Eleanor Power, Melissa Sands, Giovanna Scarchilli, Xavier Jara, and Minsu Park for valuable detailed feedback.

References

Arnould, Eric J., and Craig J. Thompson. 2005. 'Consumer Culture Theory (CCT): Twenty Years of Research'. *Journal of Consumer Research* 31(4):868–82. doi: 10.1086/426626.

Atkinson, Will. 2011. 'The Context and Genesis of Musical Tastes: Omnivorousness Debunked, Bourdieu Buttressed'. *Poetics* 39(3):169–86. doi: 10.1016/j.poetic.2011.03.002.

Attanasio, Orazio P., and Luigi Pistaferri. 2016. 'Consumption Inequality'. *Journal of Economic Perspectives* 30(2):3–28. doi: 10.1257/jep.30.2.3.

Baumann, Shyon, Michelle Szabo, and Josée Johnston. 2019. 'Understanding the Food Preferences of People of Low Socioeconomic Status'. *Journal of Consumer Culture* 19(3):316–39. doi: 10.1177/1469540517717780.

Belk, Russell W. 1988. 'Possessions and the Extended Self'. *Journal of Consumer Research* 15(2):139–68. doi: 10.1086/209154.

Berger, Jonah, and Morgan Ward. 2010. 'Subtle Signals of Inconspicuous Consumption'. *Journal of Consumer Research* 37(4):555–69. doi: 10.1086/655445.

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

Bourdieu, Pierre, and Jean-Claude Passeron. 2000. *Reproduction in Education, Society and Culture*. 2 .ed., reprinted. London: Sage Publ.

Bowles, Samuel, and Yongjin Park. 2005. 'Emulation, Inequality, and Work Hours: Was Thorsten Veblen Right?' *The Economic Journal* 115(507):F397–412. doi: 10.1111/j.1468-0297.2005.01042.x.

Chan, Tak Wing. 2019. 'Understanding Cultural Omnivores: Social and Political Attitudes'. *The British Journal of Sociology* 70(3):784–806. doi: 10.1111/1468-4446.12613.

Chao, Angela, and Juliet B. Schor. 1998. 'Empirical Tests of Status Consumption: Evidence from Women's Cosmetics'. *Journal of Economic Psychology* 19(1):107–31. doi: 10.1016/S0167-4870(97)00038-X.

Charles, Kerwin Kofi, Erik Hurst, and Nikolai Roussanov. 2009. 'Conspicuous Consumption and Race*'. *The Quarterly Journal of Economics* 124(2):425–67. doi: 10.1162/qjec.2009.124.2.425.

Christen, Markus, and Ruskin M. Morgan. 2005. 'Keeping Up With the Joneses: Analyzing the Effect of Income Inequality on Consumer Borrowing'. *Quantitative Marketing and Economics* 3(2):145–73. doi: 10.1007/s11129-005-0351-1.

Coulangeon, Philippe. 2017. 'Cultural Openness as an Emerging Form of Cultural Capital in Contemporary France'. *Cultural Sociology* 11(2):145–64. doi: 10.1177/1749975516680518.

DiMaggio, Paul. 1982. ‘Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of U.S. High School Students’. *American Sociological Review* 47(2):189–201. doi: 10.2307/2094962.

Dong, Xiaowen, Eaman Jahani, Alfredo J. Morales, Burçin Bozkaya, Bruno Lepri, and Alex ´Sandy’ Pentland. 2020. ‘Purchase Patterns, Socioeconomic Status, and Political Inclination’. *The World Bank Economic Review* 34(Supplement_1):S9–13. doi: 10.1093/wber/lhz008.

Dong, Yuxiao, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. 2014. ‘Inferring User Demographics and Social Strategies in Mobile Social Networks’. Pp. 15–24 in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’14*. New York, New York, USA: Association for Computing Machinery.

Eckhardt, Giana M., and Fleura Bardhi. 2020. ‘New Dynamics of Social Status and Distinction’. *Marketing Theory* 20(1):85–102. doi: 10.1177/1470593119856650.

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. doi: 10.1080/0267257X.2014.989890.

Goldberg, Amir. 2011. ‘Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined’. *American Journal of Sociology* 116(5):1397–1436. doi: 10.1086/657976.

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. doi: 10.1177/0003122416632787.

Goyal, Palash, and Emilio Ferrara. 2018. ‘Graph Embedding Techniques, Applications, and Performance: A Survey’. *Knowledge-Based Systems* 151:78–94. doi: 10.1016/j.knosys.2018.03.022.

Grover, Aditya, and Jure Leskovec. 2016. ‘Node2vec: Scalable Feature Learning for Networks’. Pp. 855–64 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*. New York, NY, USA: Association for Computing Machinery.

Heffetz, Ori. 2011. ‘A Test of Conspicuous Consumption: Visibility and Income Elasticities’. *The Review of Economics and Statistics* 93(4):1101–17. doi: 10.1162/REST_a_00116.

Krueger, Dirk, and Fabrizio Perri. 2006. ‘Does Income Inequality Lead to Consumption Inequality? Evidence and Theory1’. *The Review of Economic Studies* 73(1):163–93. doi: 10.1111/j.1467-937X.2006.00373.x.

Kumar, Bipul, Richard P. Bagozzi, Ajay K. Manrai, and Lalita A. Manrai. 2021. ‘Conspicuous Consumption: A Meta-Analytic Review of Its Antecedents, Consequences, and Moderators’. *Journal of Retailing*. doi: 10.1016/j.jretai.2021.10.003.

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):9. doi: 10.1007/s13278-018-0486-1.

Pechey, Rachel, and Pablo Monsivais. 2015. 'Supermarket Choice, Shopping Behavior, Socioeconomic Status, and Food Purchases'. *American Journal of Preventive Medicine* 49(6):868–77. doi: 10.1016/j.amepre.2015.04.020.

Perugini, Cristiano, Jens Hölscher, and Simon Collie. 2016. 'Inequality, Credit and Financial Crises'. *Cambridge Journal of Economics* 40(1):227–57. doi: 10.1093/cje/beu075.

Peterson, Richard A. 1992. 'Understanding Audience Segmentation: From Elite and Mass to Omnivore and Univore'. *Poetics* 21(4):243–58. doi: 10.1016/0304-422X(92)90008-Q.

Peterson, Richard A., and Roger M. Kern. 1996. 'Changing Highbrow Taste: From Snob to Omnivore'. *American Sociological Review* 61(5):900–907. doi: 10.2307/2096460.

Pons, Pascal, and Matthieu Latapy. 2006. 'Computing Communities in Large Networks Using Random Walks'. *Journal of Graph Algorithms and Applications* 10(2):191–218. doi: 10.7155/jgaa.00124.

Reimer, Suzanne, and Deborah Leslie. 2004. 'Identity, Consumption, and the Home'. *Home Cultures* 1(2):187–210. doi: 10.2752/174063104778053536.

SafeGraph. 2022. 'Open Census Data | SafeGraph Docs'. *SafeGraph*. Retrieved 12 July 2022 (https://docs.safegraph.com/docs/open-census-data).

SafeGraph. 2023. 'Welcome | SafeGraph Docs'. *SafeGraph*. Retrieved 14 January 2023 (https://docs.safegraph.com/docs/join-our-community).

Sahin, Onur, and Suphan Nasir. 2022. 'The Effects of Status Consumption and Conspicuous Consumption on Perceived Symbolic Status'. *Journal of Marketing Theory and Practice* 30(1):68–85. doi: 10.1080/10696679.2021.1888649.

Schaefers, Tobias. 2014. 'Standing out from the Crowd: Niche Product Choice as a Form of Conspicuous Consumption'. *European Journal of Marketing* 48(9/10):1805–27. doi: 10.1108/EJM-03-2013-0121.

Srivastava, Abhinav, Srabanti Mukherjee, and Charles Jebarajakirthy. 2020. 'Aspirational Consumption at the Bottom of Pyramid: A Review of Literature and Future Research Directions'. *Journal of Business Research* 110:246–59. doi: 10.1016/j.jbusres.2019.12.045.

Trigg, Andrew B. 2001. 'Veblen, Bourdieu, and Conspicuous Consumption'. *Journal of Economic Issues* 35(1):99–115. doi: 10.1080/00213624.2001.11506342.

US Census Bureau. 2022. 'American Community Survey 5-Year Data (2009-2020)'. *Census.Gov*. Retrieved 12 July 2022 (https://www.census.gov/data/developers/data-sets/acs-5year.html).

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

Veenstra, Gerry. 2015. ‘Class Position and Musical Tastes: A Sing-Off between the Cultural Omnivorism and Bourdieusian Homology Frameworks’. *Canadian Review of Sociology/Revue Canadienne de Sociologie* 52(2):134–59. doi: 10.1111/cars.12068.

de Vries, Robert, and Aaron Reeves. 2021. ‘What Does It Mean to Be a Cultural Omnivore? Conflicting Visions of Omnivorousness in Empirical Research’. *Sociological Research Online* 13607804211006109. doi: 10.1177/13607804211006109.

Walasek, Lukasz, Sudeep Bhatia, and Gordon D. A. Brown. 2018. ‘Positional Goods and the Social Rank Hypothesis: Income Inequality Affects Online Chatter about High- and Low-Status Brands on Twitter’. *Journal of Consumer Psychology* 28(1):138–48. doi: 10.1002/jcpy.1012.

Walasek, Lukasz, and Gordon D. A. Brown. 2015. ‘Income Inequality and Status Seeking: Searching for Positional Goods in Unequal U.S. States’. *Psychological Science* 26(4):527–33. doi: 10.1177/0956797614567511.

Wang, Daixin, Peng Cui, and Wenwu Zhu. 2016. ‘Structural Deep Network Embedding’. Pp. 1225–34 in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*. New York, NY, USA: Association for Computing Machinery.

Warde, Alan. 2015. ‘The Sociology of Consumption: Its Recent Development’. *Annual Review of Sociology* 41(1):117–34. doi: 10.1146/annurev-soc-071913-043208.

Warde, Alan, David Wright, and Modesto Gayo-Cal. 2008. ‘The Omnivorous Orientation in the UK’. *Poetics* 36(2):148–65. doi: 10.1016/j.poetic.2008.02.004.

Wisman, Jon D. 2009. ‘Household Saving, Class Identity, and Conspicuous Consumption’. *Journal of Economic Issues* 43(1):89–114.

Woodward, Ian. 2003. ‘Divergent Narratives in the Imagining of the Home amongst Middle-Class Consumers: Aesthetics, Comfort and the Symbolic Boundaries of Self and Home’. *Journal of Sociology* 39(4):391–412. doi: 10.1177/0004869003394005.

Zagorsky, Jay L., and Patricia K. Smith. 2017. ‘The Association between Socioeconomic Status and Adult Fast-Food Consumption in the U.S’. *Economics and Human Biology* 27(Pt A):12–25. doi: 10.1016/j.ehb.2017.04.004.

Chapter 5 Socioeconomic Inequality in Social Capital and Communication Behaviour on Twitter

Abstract

The pervasiveness of socioeconomic inequality could extend into social media platforms like Twitter. However, relevant empirical evidence remains rare and fragmented. Leveraging a recently developed method for estimating Twitter users' individual socioeconomic status (SES), this study investigates socioeconomic inequality in social capital and communication behaviours on Twitter. First, this paper establishes that higher SES Twitter users have higher social capital across different measures of social capital, continuing the recent efforts in quantifying the relationship between socioeconomic outcomes and social capital in large digital networks. Second, compared with the existing scattered evidence, this paper provides a more comprehensive picture of the relationship between SES and communication behaviours. The paper demonstrates that higher SES users use more complex and future-oriented language in their tweets. Also, while high and low SES users mostly talk about similar topics, they tend to use different hashtags and have divergent sentiments towards immigration. These findings reveal that socioeconomic inequalities are not only reflected but also potentially reinforced on social media, underscoring the critical roles of social capital and communication behaviours. The study highlights the need for further research to explore the underlying mechanisms and integrate SES as a critical factor in social media research.

Introduction

Socioeconomic inequality is deeply intertwined with daily behaviour and social interactions, where it is continuously manifested, maintained, and reproduced. Individuals display their socioeconomic status (SES) through verbal or nonverbal cues, compare their SES with others, and encounter socioeconomic disparities in their everyday lives. This constant exposure to socioeconomic inequality could contribute to the maintenance and reproduction of inequality (Bourdieu 1984; Kraus, Park, and Tan 2017). As social media platforms like Twitter have become integral to daily lives and social interactions, it is crucial to explore how socioeconomic inequality is reflected and reinforced on these platforms.

However, research in this area is hampered due to the limited availability of individual-level SES information on social media platforms. There are few existing papers on the relationship between SES and social capital or communication behaviours. On the one hand, existing literature on the relationship between SES and social capital in large digital communication networks is relatively nascent, with a few notable studies at the community and individual levels (Chetty et al. 2022; Eagle, Macy, and Claxton 2010; Luo et al. 2017; Norbutas and Corten 2018). These studies use telephone communication networks or social media platforms as proxies for real-life networks, so they neglect what happens on social media platforms per se. On the other hand, the literature on the relationship between SES and communication behaviours mainly focuses on using communication patterns to estimate SES rather than examine the relationship through a perspective of inequality (Abitbol, Fleury, and Karsai 2019;

Preoţiuc-Pietro, Volkova, et al. 2015). Consequently, the findings remain fragmented and difficult to generalise.

This paper addresses these gaps by presenting the first systematic investigation of socioeconomic inequality in social capital and communication behaviours on Twitter. Since the data were collected before Twitter was renamed X, I will refer to the platform as Twitter throughout this paper. This paper focuses on Twitter due to its crucial role in society and widespread use in academic research (Murthy 2024). Despite the popularity of Twitter in academic research, there is a significant gap in studying socioeconomic inequality on Twitter. SES is not even accounted for as a control variable in most existing literature, which could be problematic given its potential associations with popular Twitter research areas such as health, politics and social movements.

Leveraging the recently developed method for estimating individual SES of Twitter by He and Tsvetkova (2023), this paper systematically analyses the socioeconomic inequality in social capital and communication behaviour on Twitter. On the one hand, the paper studies socioeconomic inequality in social capital on Twitter by testing the hypotheses on the relationship between SES and social capital across different measures of social capital. On the other hand, the paper explores the socioeconomic inequality in communication behaviour on Twitter through three angles: hypothesis testing, descriptive exploration of topics, and a case study. First, the paper tests the hypotheses that Twitter users with higher SES tend to use more complex and future-oriented language in their tweets. Second, the paper provides a descriptive analysis of the topics discussed by high- versus low-SES users. Third, the paper presents a case study of the relationship between sentiments towards immigration and SES. Overall, the paper illustrates that Twitter users with higher SES tend to have higher social capital and more advantageous communication behaviours. The findings encourage further research in the area and highlight the importance of accounting for SES when researching social media platforms.

Socioeconomic status and social capital

Social capital is a concept that captures the value embedded within individuals' social networks, often determining access to resources, information, and opportunities. However, social capital operates at multiple levels and is measured through various dimensions, leading to inconsistent findings regarding its relationship with socioeconomic outcomes (Portes 1998; Westlund and Adam 2010). A useful theoretical distinction of social capital, particularly for the scope of this paper, is between competitive and cooperative social capital (Gelderblom 2018; Julien 2015; Lin 1999; Portes 1998). This distinction allows for a more nuanced understanding of how social capital might interact with socioeconomic status (SES) in both physical and digital environments.

Competitive social capital refers to forms of social capital that provide individuals with competitive advantages through the size, strength, and structure of their social networks. This concept encompasses Granovetter's (1973) theory of weak ties, Bourdieu's (1986) view of social capital's instrumental value to individuals, Burt's (1992) theory of structure holes, and Putnam's (2000) idea of bridging capital. In contrast, Cooperative social capital describes the kind of social capital that people enjoy in a tightly connected community, where a high level of trust facilitates activities beneficial to the socioeconomic prosperity of everyone in the community. This type of social capital aligns with Coleman's (1988) focus on trust from network closure and Putnam's (2000) emphasis on civic engagement. Operationally, competitive social capital is better analysed at the individual level but can be aggregated to the

group level. On the contrary, cooperative social capital is typically examined at the community level and is hard to disaggregate to the individual level.

The relationship between socioeconomic status and social capital is well-established in the literature. On the one hand, social capital can affect socioeconomic outcomes by facilitating the information flow, behaviour adoption, and social cohesion beneficial to SES (DiMaggio and Garip 2012; Granovetter 2005; Jackson 2021). On the other hand, SES may influence access to social capital and the ability to leverage it (Lin 2000; Mouw 2003). These bidirectional dynamics illustrate the crucial role of social capital in reinforcing socioeconomic inequality.

However, research on the relationship between socioeconomic status (SES) and social capital on large digital communication networks is constrained due to the limited availability of socioeconomic data. Even studies at the geographical unit level are relatively rare in this domain: notable research includes studies by Eagle, Macy, and Claxton (2010) and Norbutas and Corten (2018). The former links the UK mobile phone communication network with economic development at the level of telephone regional exchange areas, while the latter connects the data of Hyves (a Dutch social media platform) and socioeconomic measures at the level of municipalities. At the individual level, only two studies have been identified. Luo et al. (2017) link telephone communication networks with bank credit data in Mexico. Chetty et al. (2022) connect a large sample of US Facebook users with representative survey data.

On the one hand, existing studies reveal positive associations between SES and competitive social capital across different measures. Eagle et al. (2010) illustrate the positive associations between economic development and competitive social capital at the geographical unit level in the UK with measures including topological diversity, spatial diversity, structural holes (effective size), and degree centrality. Norbutas and Corten (2018) replicate the findings of topological diversity and spatial diversity in Dutch communities. They also show nuances of the relationship between topological diversity and economic development depending on the measures of economic development and controls. At the individual level, Luo et al. (2017) replicate the positive associations with network diversity and centrality measures. While extending the finding of degree centrality to more complex centrality measures (PageRank centrality, k-cell index and Collective Influence Metric), they use a simpler network diversity than topological diversity—the ratio of communication within to outside individuals' community. Chetty et al. (2022) show that economic connectedness, measured as the proportion of high-SES friends among people with low SES, strongly correlates with upward income mobility.

On the other hand, the results of cooperative social capital are inconclusive from existing literature. Only Chetty et al. (2022) and Norbutas and Corten (2018) included cooperative social capital measures, and the associations between socioeconomic outcomes and cooperative social capital vary depending on the measures used. This inconsistency highlights the difficulty of measuring cooperative social capital. Cooperative social capital is primarily measured at the community level, but the relationship between cooperative social capital and socioeconomic outcomes is sensitive to the size of the community (Westlund and Adam 2010). It is possible that the size of the communities in Chetty et al. (2022) and Norbutas and Corten (2018) is not where cooperative social capital operates.

This paper advances the study of the relationship between socioeconomic status (SES) and social capital in large digital communication networks by conducting individual-level analysis

with Twitter data. Existing studies treat digital communication networks as proxies for offline networks. Chetty et al. (2022) and Norbutas and Corten (2018) explicitly stated in their papers that they treat the friendship networks in social media platforms as proxies for real-life networks, while telephone networks in nature are a proxy for offline networks and cannot be a social space on their own. In contrast, this paper makes a unique contribution by studying Twitter's online social interactions per se. A distinct feature of Twitter's online communication is the @mentions. As interactions indicate a minimal level of reciprocity, this paper focuses on the networks constructed by users with reciprocal @mentions. Therefore, this paper examines the relationship between individual users' SES and their social capital that emerged from the reciprocal @mentions networks on Twitter.

An overarching argument of this paper is that Twitter users with higher SES tend to have higher social capital. Specifically, this paper tests the hypotheses on the relationship between SES and six measures of social capital that are chosen based on existing literature and data availability.

The first measure is degree centrality—the number of contacts of a user (for convenience, I use the term contacts to refer to users with reciprocal @mentions). Degree centrality provides a basic understanding of the users' social capital. Therefore, the first specific hypothesis of this paper is:

H1. *Twitter users with higher SES have higher degree centrality.*

The second measure is reciprocity—the proportion of @mentions the users send out that are reciprocated. Reciprocity reflects the quality and strength of user interactions in their networks. Therefore:

H2. *Twitter users with higher SES have higher reciprocity.*

The third measure is topological diversity, as in Eagle et al. (2010) and Norbutas and Corten (2018). Topological diversity measures the diversity of the users' interactions. So, to replicate existing studies:

H3. *Twitter users with higher SES have higher topological diversity.*

The fourth measure is local clustering coefficient, the proportion of the number of actual ties between a user's contacts divided by the number of possible ties between the user's contacts. Local clustering coefficient is used as a proxy for the possibility of weak ties in a user's ego network. Higher local clustering means a lower chance of weak ties and, thus, lower social capital. Therefore:

H4. *Twitter users with higher SES have a lower local clustering coefficient.*

The fifth and sixth measures, effective size and efficiency, are based on Burt's (1992) structure holes theory. Effective size measures to what extent a user's contacts are not redundant; thus, the user occupies an advantageous position in the information flow. Efficiency is defined as the effective size divided by degree. While effective size indicates the breadth of unique information accessible to the user, efficiency reflects how optimised the users are on obtaining unique information per contact. Therefore

H5. *Twitter users with higher SES have a larger effective size.*

H6. *Twitter users with higher SES have higher efficiency.*

Socioeconomic status and communication behaviours

Like social capital, the relationship between SES and communication behaviours is bidirectional. SES can affect individuals' communication behaviours, which in turn may result in beneficial socioeconomic outcomes (Bernstein 1960; Hazen and Black 1989; Manstead 2018). However, the limited availability of individual-level socioeconomic data again constrains research on this relationship in social media platforms like Twitter. Most research on the topic tends to focus on using communication patterns to estimate Twitter users' SES, where communication patterns tend to be one of the many features used for SES estimation with a complex machine learning model. The relationship between SES and communication behaviours is often not systematically examined: Some research does not show the contribution of communication behaviours in their SES estimation (Lampos et al. 2016), while others show scattered results that are hard to generalise (Abitbol et al. 2019; Preoţiuc-Pietro, Volkova, et al. 2015). One rough pattern from the scatter results is that high SES Twitter users tend to talk more about politics, technology, business, art and literature, while low SES users are more likely to talk about beauty care and use informal or cursing language (Abitbol and Morales 2021; Preoţiuc-Pietro, Volkova, et al. 2015; Preoţiuc-Pietro, Lampos, and Aletras 2015).

This paper systematically analyses the relationship between SES and communication behaviours on Twitter. This paper approaches the issue from three angles: hypothesis testing, descriptive exploration of topics, and a case study.

This paper starts with the very basics by hypothesis testing the relationship between SES and simple measures of communication behaviours. One of the most essential features of communication behaviours is the complexity of the language used. Based on existing research on the positive associations between SES and language complexity (Bradac et al. 1977; Flekova, Preoţiuc-Pietro, and Ungar 2016; Sankoff and Lessard 1975), it is reasonable to expect:

H7. *Twitter users with higher SES use more complex language in their tweets.*

Another simple communication behaviour related to SES is future orientation. For example, Preis et al. (2012) showed that people in wealthier countries tend to search more future years on Google. Ireland et al. (2015) illustrated that US counties with higher rates of future tense tweets tend to have fewer HIV cases. This paper hypothesises that:

H8. *Twitter users with higher SES use more future-oriented language in their tweets.*

Next, this paper compares topics tweeted by users from high vs low SES backgrounds. I first compare the most frequently used hashtags by users with different SES. Then, I use topic modelling to identify the topics tweeted by users in the sample and test which topics have significantly divergent probabilities of being tweeted by users from high or low SES backgrounds.

Finally, this paper conducts a case study of the relationship between SES and sentiments towards immigration. I extract tweets about immigration from the sample and test whether sentiments towards immigration significantly differ by SES. This paper analyses three aspects of sentiments towards immigration, estimated by relevant large language models. The first aspect is simply positive-negative sentiment, estimating the probabilities of the tweets being positive, negative, and neutral. The second aspect is about hate speech, detecting the probabilities of the tweets being hateful, targeted, and aggressive. The third aspect is sentiments toward authority. Based on the Moral Foundations Theory (Graham et al. 2013; Graham, Haidt, and Nosek 2009), I estimate the probabilities of the tweets that submit to authority and tradition (authority) and reject the subversion of authority and tradition (subversion).

Methods

Data

Based on the distribution of estimated SES of Twitter users from He and Tsvetkova (2023), I selected the users who are at least three standard deviations lower or 1.5 standard deviations higher than the mean, resulting in 21,908 users at the higher end and 21,209 users at the lower end. Then, I used the latest version of a well-established bot detection method, Botometer, and a recommended threshold of 0.75 to filter out potential bots (Martini et al. 2021; Sayyadiharikandeh et al. 2020). After the bot detection, 16,467 users at the higher end and 11,742 users at the lower end remained. For the 28,209 remaining users in the sample, I used the Academic Track of Twitter API to collect their user objects (account metadata), the user ID of all their followers and accounts they follow, all their tweets up to the latest 3200, and all the tweets that mentioned them up to the latest 800. Due to reasons including private, deleted, and suspended accounts, I ended up with data for 27,445 users, including 15,839 with high SES and 11,606 with low SES. For the analyses about users' tweets in this paper, I selected only English tweets that are not retweets, resulting in 33,883,064 tweets, where 16,923,693 are from high SES users and 16,959,371 are from low SES users.

From the users' tweets and mentions, I identified their contacts, which are defined as the accounts with whom they have mutual @mentions. Then I tried to collect all the mentions since 2019 between each user's contacts, if they exist. Before the closure of Twitter Academic API in May 2023, I was able to obtain the tweets between contacts for 7,674 users, including 4,308 with high SES and 3,366 with low SES.

Matching

This paper used matching to compare the relevant quantitative measures between high and low SES users. Matching was conducted following Ho et al.'s (2007) nonparametric pre-processing approach and its R implementation, "MatchIt" (Ho et al. 2011). I used high-SES users as the treatment group and low-SES users as the control group to estimate the 'average treatment effect (ATE)' of being in high SES compared with low SES on social capital, language complexity, and other interested measures in this paper. Because this paper aims to identify associations instead of causal effects, the process effectively compares the interested variables between the two SES groups, controlling for the potential confounding variables. The 'ATE' in this case means the average differences of the interested measures between users in the two SES groups.

The matching was conducted in the following steps. First, I identified the potential confounders as the covariates for matching. Eight potential confounders were available from the data: the number of follows, the number of followers, the number of mutual follows, the number of lists the user is a member of, the number of likes, the number of tweets, the time passed since the accounted was registered in seconds (I chose the end time to be when the newest account in the sample is registered), and whether the account is verified. As degree centrality by contacts is the most basic social capital measure and a potential confounder for other measures in this study, I also included degree centrality as a covariate for matching when it was not the interested measure.

Second, I attempted different matching methods and selected the best-performing one based on balance tests. I tried Generalised Full Matching, Optimal Full Matching, Coarsened Exact Matching, and Subclassification. Generalised Full Matching with Mahalanobis distance turned out to perform the best whether degree centrality is included as a covariate and for various subsets of the sample. Decent balances were achieved for all cases. In most cases, absolute standardised mean differences for all the covariates were below 0.1 after matching. Even in the most restricted cases, only one covariate had absolute standardised mean differences slightly higher than 0.1 but still lower than 0.15. Generalised Full Matching uses all treated and all control units, so no units were discarded. Detailed matching results are available in Appendix A.

Third, for each interested measure, to estimate its average differences between the two SES groups, standard error and significance, I fit a linear regression model with the measure as the outcome and the SES group, matching covariates, and their interaction as predictors, and included the full matching weights in the estimation. Then, I used the R package "marginaleffects" to perform g-computation in the matched sample to estimate the average difference. A cluster-robust variance was used to estimate its standard error with matching stratum membership as the clustering variable.

Social Capital Measures

For Twitter users in the sample, I construct ego networks where the sampled users are the egos, and their contacts are the alters. Then, I measure the users' social capital by six different network properties in their ego networks. The first three measures utilise radius-1 ego networks: they only use the @mentions between the users and their contacts. The last three measures add information from the radius-2 ego networks: they also concern the @mentions between each user's contacts. Therefore, radius-1 measures are available for all 27,445 users in our sample, whereas radius-2 measures are only available for the 7,674 users whose tweets between their contacts are available.

The measures are calculated in two ways to deal with data limitation and provide a robustness test. Collecting the users' tweets and the tweets mentioned the users from the Twitter API are bounded by the 3200 and 800 upper limits, while collecting the tweets between users' contacts does not. Therefore, for users with more than 3200 tweets or 800 incoming mentions, the number of interactions between users and contacts may be incomplete compared with the number of interactions between contacts. The imbalance of the number of interactions in the network means unbalanced weights for the edges, so the radius-2 measures that account for weights might be problematic.

To deal with the weight imbalance, I conduct two sets of calculations for the social capital measures, and both sets are used for later analyses to complement each other. In the first set, I use all the data available to construct the ego networks and calculate the radius-2 measures by treating the ego networks as unweighted and undirected networks. In the second set, I construct the ego networks with only @mentions that happened in 2022 and calculate radius-2 measures accounting for weights. The first set emphasises edge existence structure, whereas the second set emphasises weights. The two sets provide reasonable complements with each other. The imbalance issue does not affect radius-1 measures, so the calculations for radius-1 measures remain the same. Nevertheless, radius-1 measures from more limited data in the second set provide a useful robustness test.

The radius-1 measures used in this study are degree centrality, reciprocity, and topological diversity. Degree centrality means the number of contacts the users have. It measures the size of the ego network and provides a basic understanding of the users' social capital. Reciprocity is the proportion of @mentions the users send out that are reciprocated. It reflects the quality and strength of the interactions users enjoy in their ego networks. Topological diversity, as used by Eagle, Macy, and Claxton (2010), is measured as the normalised Shannon entropy. For a user i, the topological diversity is :

$$D(i) = \frac{-\sum_{j=1}^{k} p_{ij} \log (p_{ij})}{\log (k)}$$

where k is the number of i's contacts (degree centrality) and p_{ij} is the proportion of @mentions between i and a contact j out of all @mentions between i and all i's contacts. Topological diversity measures the diversity of the users' interactions with their contacts. It will be low if the users mainly interact with a few contacts and high if the users interact with their contacts more evenly.

The radius-2 measures used in this study are local clustering coefficient, effective size, and efficiency. Local clustering coefficient is defined as the proportion of the number of actual ties between a user's contacts divided by the number of possible ties between the user's contacts. For unweighted and undirected ego networks, the local clustering coefficient of a user i is:

$$C(i) = \frac{2T(i)}{k(k-1)}$$

where $T(i)$ is the number of ties between i's contacts, and k is the number of i's contacts (Watts and Strogatz 1998). For weighted and directed ego networks, the local clustering coefficient of a user i is:

$$C(i) = \frac{T(i)}{2[k_i^{tot}(k_i^{tot} - 1) - 2k_i^{\leftrightarrow}]}$$

where $T(i)$ is the number of directed triangles though i, k_i^{tot} is the sum of in degree and out degree of i and $k_i^{\leftrightarrow}$ is the reciprocal degree of i (Fagiolo 2007). Local clustering coefficient is used as a proxy for the possibility of weak ties in a user's ego network. Higher local clustering means a lower chance of weak ties and, thus, lower social capital.

Effective size measures to what extent a user's contacts are not redundant and thus occupy an advantageous position in the information flow (Burt 1992). The effective size for a user i is:

$$E(i) = \sum_j \left[1 - \sum_q p_{iq} m_{jq}\right], q \neq i, j$$

where j is one of i's contacts, q is another contact, p_{iq} is the normalised mutual weight of the ties between i and q, and m_{jq} is the mutual weight of j and q divided by j's highest mutual weight with any of its contacts. The mutual weight between two nodes is the sum of the weights between them. For unweighted and undirected ego networks, the effective size is measured using a simplified formula by Borgatti (1997):

$$E(i) = k - \frac{2T(i)}{k}$$

where $T(i)$ is the number of ties between i's contacts, and k is the number of i's contacts.

Efficiency is defined as the effective size divided by degree. While effective size indicates the breadth of unique information accessible to the user, efficiency reflects how optimised the users are on obtaining unique information per contact.

The correlations between the six social capital measures for the full and 2022 samples are shown in Tables B1 and B2 in Appendix B. Not surprisingly, most of the measures have weak associations with each other. Degree centrality is strongly associated with effective size but weakly associated with efficiency, indicating the utility of including efficiency in the study.

Language complexity Measures

Language complexity is measured by lexical diversity and readability scores. Lexical diversity measures the diversity of the type of words used in the text. The simplest measure of lexical diversity is the Type-token Ratio (TTR), which is the total number of token types divided by the total number of tokens. However, TTR is too sensitive to text length and has limited usefulness when comparing texts with different lengths (Bestgen 2024). Therefore, I use three measures of lexical diversity that address that issue in different ways. The first measure is the Hypergeometric Distribution-Derived (HD-D) index, which accounts for the probability distribution of token types in the text (McCarthy and Jarvis 2007). The second measure is the Mean Segmental Type-Token Ratio (MSTTR), which divides the text into contiguous segments and calculates the average TTR of the segments (Malvern et al. 2004). The third measure is the Measure of Textual Lexical Diversity (MTLD), which is the mean length of sequential token strings in a text that maintains a particular TTR value (McCarthy and Jarvis 2010).

Readability measures how easily a text can be understood. I use the Flesch–Kincaid Grade Level test, which scales the score to the US grade level(Kincaid et al. 1975). The score is a function of words, syllables and sentences:

$$Readibility = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$$

To calculate the language complexity measures, I pre-processed the tweets by removing the URLs, mentions, hash symbols, numbers, and emojis and converting all text to lowercase. Following this pre-processing, I employed the Python library "textstat" to calculate the Flesch–Kincaid Grade Level. Subsequently, I utilised the "lexical-diversity" Python library to lemmatise the pre-processed tweets and calculate the three lexical diversity measures.

The correlations between the language complexity measures are shown in Table B3 in Appendix B. Unsurprisingly, the lexical diversity measures are strongly associated with each other and are weakly associated with the Readability measure.

Future Orientation

The level of future orientation is measured by the proportion of sentences of a user's tweets that use future tense words: *will, would, shall, going to, gonna, won't.* This measure is similar to Ireland et al. (2015), but more conservative in the choice of future tense words. Compared with Ireland et al. (2015), I excluded modal verbs like *may, might, must*, etc. to only include words with a high probability of indicating future tense. Like language complexity measures, the pre-processing of the tweets for the future orientation measure involved removing URLs, mentions, hash symbols, numbers, and emojis and converting all text to lowercase.

Hashtags

I used regular expressions to extract hashtags from the tweets in the sample. Of 34,698,277 tweets in our sample, 11,363,176 (32.7%) have hashtags. The most frequent hashtags that contain only numbers do not represent meaningful topics, such as '#1' or '#2'. Therefore, I excluded the hashtags that contained only numbers, resulting in 11,301,356 (32.7%) tweets with 1,182,413 not purely numerical hashtags. For descriptive comparisons, I ranked the hashtags by the number of users who used the hashtags and plotted word clouds with the most frequent 100 hashtags for high and low SES users.

Topic modelling

This paper focuses on a general understanding of the topics at the user level instead of the tweet level, so I treat all tweets of a user as one document for topic modelling. I used the Python library "genism" to conduct the Latent Dirichlet Allocation (LDA) approach (Blei, Ng, and Jordan 2003). I implemented the recommended best practices from Laureate, Buntine, and Linger's (2023) systematic review. I pre-processed the tweets by removing URLs, mentions, hash symbols, numbers, emojis, Unicode characters, stop words, punctuations and pure numbers, converting contractions, and lemmatising the tokens. Then, I filtered tokens by removing those that appeared in below 20 documents or more than half of the documents.

To select the hyperparameters, I first run models with iteration = 400, pass = 50, and k (the number of topics) = {10, 50, 100}. The convergence score stabilised around passes 30-40 for the three models, indicating iteration = 400 and pass = 50 a decent combination. I inspected the topic clusters and related most frequent words when k = 50 and found overlapping topics, so the number of topics should be less than 50. Then, I run models with k = {20, 30, 40}. After inspecting the topic clusters, it becomes clear that the optimal topic is around 30. So, I next run models with k from 25 to 39. I tried to use coherence and perplexity scores to choose the optimal k, but the scores turned out to be too close to be informational. Therefore, I inspected the topic clusters of k from 25 to 40 and found that the clusters make the best sense when setting k to 30. I obtained the probabilities of the users tweeting about the 30 topics and used matching to compare the probabilities of tweeting about the topics between high and low SES users.

Although k = 30 makes the most sense, some of the 30 topics can be merged into larger groups. For example, the four topics, 'elections', 'international_affairs', 'party_politics', and 'party_politics_emotions', can be merged into one large group (I named it 'merged_politics'). I identified four such groups and merged the topics by taking the average probabilities of the topics in the groups. I also added the four large topic groups to the matching comparison of the topics. The top ten words associated with the 30 topics and the topics associated with the four larger groups are available in Appendix C.

Case study of sentiments towards immigration

To compare the sentiments towards immigration between high and low SES users, I first extracted tweets about immigration in the sample. I selected relevant words and hashtags such as 'immigrant', 'immigration', 'undocumented worker', 'openborder', 'keepthemout' based on previous immigration research with Twitter (Menshikova and van Tubergen 2022; Rowe et al. 2021). Some of the words are neutral (e.g., 'immigrant'), while others already imply positive

or negative sentiments(e.g., ‘openborder’ is positive, ‘keepthemout’ is negative). The complete list is available in Appendix D. I also excluded the tweets about the Immigrant Song. Then, I used text matching to find 72,232 tweets about immigration in the sample. There are 68,176 tweets from 9000 high SES users and 3941 tweets from 1470 low SES users.

Next, I used finetuned-BERT models to estimate the different aspects of sentiments. The models were chosen because they were shown to be the best-performing approach for these tasks among dictionary methods, supervised learning models, and zero-shot classification with GPT4 (Macanovic and Przepiorka 2024). I employed the Python library “pysentimiento” for sentiment analysis and hate speech detection because the relevant models in the library were finetuned with Twitter datasets (Pérez et al. 2024). Macanovic and Przepiorka (2024) illustrated that a finetuned RoBERTa model performs best in estimating moral sentiments, including authority and subversion with the Moral Foundations Twitter Corpus (Hoover et al. 2020). Therefore, I used the same code and data as Macanovic and Przepiorka (2024) to build an authoritarian sentiment estimation model.

After estimating the sentiment measures, I aggregated the measures to the user level by taking the average and used matching to estimate the differences in the sentiment measures of immigration tweets between high and low SES users. As a baseline reference, I also estimated the sentiment differences of randomly sampled tweets in the sample. The random sampled tweets consist of 100,000 tweets from 14,000 high SES users and 100,000 tweets from 10,998 low SES users. Matching for the immigration and baseline samples achieved a decent balance, where absolute standardised mean differences for all the covariates were below 0.1. Generalised Full Matching with propensity score estimated by logistic regression achieved the best balance for the immigration subsample, while Generalised Full Matching with Mahalanobis distance achieved the best balance for the baseline subsample. The detailed matching results for these two samples are available in Appendix E.

Results

Hypothesis Testing

Table 1 shows the average difference after matching between high and low SES users for social capital measures. Confirming hypotheses H1 to H6, high SES Twitter users tend to have higher social capital across all measures. High SES Twitter users have a significantly higher degree centrality than low SES users. Controlling for degree centrality, High SES Twitter users have significantly lower local clustering coefficients and higher reciprocity, topological diversity, effective size, and efficiency than low SES users. The results for the more restrictive sample constructed by @mentions in 2022 show the same pattern; the detailed results are available in Appendix F.

Table 1. Average differences after matching between high and low SES users for social capital measures.

	Estimate	Standard Error	P	2.50%	97.50%
Degree centrality	9.44	0.654	<0.001	8.15	10.7
Reciprocity	0.0419	0.00355	<0.001	0.0349	0.0489
Topological diversity	0.039	0.00213	<0.001	0.0348	0.0431
Local clustering coefficient	-0.0121	0.00267	<0.001	-0.0174	-0.0069
Effective size	0.496	0.0884	<0.001	0.323	0.67
Efficiency	0.0123	0.00228	<0.001	0.00782	0.0168

Table 2 shows the average differences after matching between high and low SES users for communication pattern measures. Confirming hypotheses H7 and H8, higher SES users tend to use more complex and future-oriented language in their tweets than low SES users.

Table 2. Average differences after matching between high and low SES users for communication pattern measures.

	Estimate	Standard Error	P	2.50%	97.50%
Readability	1.07	0.0293	<0.001	1.01	1.12
HD-D (lexical diversity)	0.0255	0.00064	<0.001	0.0243	0.0268
MSTTR (lexical diversity)	0.0263	0.000572	<0.001	0.0251	0.0274
MTLD (lexical diversity)	51.6	0.761	<0.001	50.1	53.1
Future Orientation	0.00907	0.000562	<0.001	0.00797	0.0102

Topic Comparison

Figure 1 presents word clouds of the 100 most frequent hashtags used by high and low SES users. It appears that high SES users tend to use hashtags related to COVID-19, politics and entertainment, whereas low SES users predominantly use hashtags about promotions. It makes sense that low SES users may copy and paste promotional tweets to receive discounts for certain products. However, the dominance of promotional hashtags by low SES users may raise potential concerns. As the SES is estimated based on the brands the users follow, there is a possibility that the estimated low SES users include managed accounts that exist solely to send promotional tweets. A qualitative exploration of tweets from low SES users in the sample suggests that authentic users do use promotional hashtags, and the best available way to distinguish authentic users from potentially managed accounts is the frequency of the promotional tweets in their timelines.

To address the issue, I selected a subsample that does not contain promotional tweets and accounts by removing tweets containing promotional words and hashtags and excluding users for whom more than 60 per cent of their tweets are promotional. The promotional words and hashtags are available in Appendix G. As a robustness check, I conducted the analyses for hypothesis testing again with the subsample without promotion tweets and users. The results are the same, illustrating the robustness of our hypotheses. The detailed results are available in Appendix H.

Figure 2 shows word clouds of the 100 most frequent hashtags used by high and low SES users after removing promotional tweets and users. Both high and low SES users use entertainment-related hashtags like '#superbowl' and '#music' and Covid-19-related hashtags. Apart from those, high SES users mainly use hashtags related to politics, whereas low SES users mainly use hashtags about holidays. It is worth noting that while hashtags about politics are salient among the 100 most frequent hashtags by high SES users, they are almost non-existent among the 100 most frequent hashtags by low SES users.

Figure 1. Word clouds of the 100 most frequent hashtags used by high SES and low SES users.

Figure 2. Word clouds of the 100 most frequent hashtags used by high SES and low SES users, excluding promotional tweets and users.

Table 3 shows the average differences after matching between high and low SES users for the probabilities of tweeting about the LDA-identified topics. There are no significant differences for most of the topics. High SES users are less likely than low SES users to tweet about the topic "happy_houselife" (ten most frequent words: snow, lovely, omg, husband, chocolate, yay, cheese, fantastic, cream, cake). They are also marginally less likely than low SES users to talk about the topic "closet"(ten most frequent words: added, poshmark, shopmycloset, closet, wordle, playlist, fashion, item, 4/6, listing). It is worth noting that while high SES users are more likely to use hashtags related to politics than low SES users, the two groups do not have

significant differences in the probability of tweeting about the politics-related topics generated from the topic modelling.

Table 3. Average differences after matching between high and low SES users for the probabilities of tweeting about the LDA-identified topics.

	Estimate	Standard Error	P	2.50%	97.50%
international_affairs	-0.0017	0.0011	0.123	-0.0039	5.00E-04
streaming_gaming	-0.0013	0.0013	0.3314	-0.0038	0.0013
covid19	7.00E-04	9.00E-04	0.3896	-9.00E-04	0.0024
instantwingame	1.00E-04	5.00E-04	0.9093	-9.00E-04	0.001
giveaway_pc_gaming	0.002	0.0012	0.0975	-4.00E-04	0.0043
tech_industry	0.0013	0.0014	0.3726	-0.0015	0.0041
food_promotions	0.0012	0.0012	0.3379	-0.0012	0.0036
giveaway	0.0011	0.0022	0.6026	-0.0032	0.0054
elections	-8.00E-04	0.0019	0.6605	-0.0046	0.0029
closet	-9.00E-04*	5.00E-04	0.04	-0.0018	0
party_politics_emotions	9.00E-04	0.0022	0.6931	-0.0035	0.0052
diy	-7.00E-04	7.00E-04	0.3081	-0.0021	7.00E-04
discourse_markers	-7.00E-04	0.0018	0.6899	-0.0042	0.0028
retweet_ipad_kindle	-2.00E-04	7.00E-04	0.7552	-0.0015	0.0011
entertainment	-0.0011	0.0013	0.4035	-0.0036	0.0015
strong_emotions	0.001	0.0018	0.586	-0.0026	0.0046
education_research	1.00E-04	0.0014	0.9625	-0.0026	0.0027
conference	3.00E-04	0.001	0.755	-0.0016	0.0022
sports	2.00E-04	0.0015	0.8711	-0.0028	0.0033
sweepstakes	0.0049	0.0026	0.0572	-2.00E-04	0.01
party_politics	0.002	0.0016	0.2083	-0.0011	0.0051
giveaway_scarf_necklace	5.00E-04	0.0016	0.7719	-0.0026	0.0036
cities	-5.00E-04	4.00E-04	0.2235	-0.0012	3.00E-04
reward_action	-2.00E-04	5.00E-04	0.6006	-0.0012	7.00E-04
house_construction	-2.00E-04	0.0011	0.8284	-0.0025	0.002
gratitude	-9.00E-04	0.0011	0.4394	-0.003	0.0013
nyc_ca_chicago	-6.00E-04	0.0011	0.5776	-0.0027	0.0015
climate	4.00E-04	9.00E-04	0.6933	-0.0015	0.0022
happy_houselife	-0.0073***	0.0019	1.00E-04	-0.0111	-0.0036
sweepstakes_a1_a5	4.00E-04	0.001	0.6629	-0.0015	0.0024
merged_promotions	0.0032	0.0017	0.0569	-1.00E-04	0.0066
merged_politics	-1.00E-04	0.0016	0.9368	-0.0033	0.003
merged_cities	-8.00E-04	0.001	0.3992	-0.0027	0.0011
merged_entertainment	-0.0013	0.0016	0.4285	-0.0045	0.0019

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests).

Case study of sentiments towards immigration

Tables 4 and 5 show the average differences in sentiment after matching between high and low SES users in general and towards immigration, respectively. In general, high SES users' tweets tend to be significantly less positive and more negative, hateful, targeted, aggressive and authoritarian. However, the trend is notably reversed when tweeting about immigration. In the context of immigration, high SES users' tweets tend to be significantly more neutral and less hateful, targeted, aggressive and authoritarian.

Table 4. Average differences in sentiments after matching between high and low SES users in general.

		Estimate	Standard Error	P	2.50%	97.50%
Sentiment Analysis	**Negative**	0.143***	0.0037	<0.001	0.1358	0.1502
	Neutral	-0.0035	0.0037	0.34	-0.0108	0.0037
	Positive	-0.1394***	0.0043	<0.001	-0.1479	-0.131
Hate speech detection	**Hateful**	0.0134***	0.001	<0.001	0.0114	0.0154
	Targeted	0.0037***	7.00E-04	<0.001	0.0022	0.0051
	Aggressive	0.0046***	5.00E-04	<0.001	0.0036	0.0055
Sentiment toward authority	**Authority/Subversion**	0.0083***	3.00E-04	<0.001	0.0077	0.0088
	Authority	0.0064***	2.00E-04	<0.001	0.0059	0.0069
	Subversion	0.0101***	4.00E-04	<0.001	0.0093	0.0109

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests).

Table 5. Average differences in sentiments after matching between high and low SES users towards immigration.

		Estimate	Standard Error	P	2.50%	97.50%
Sentiment Analysis	**Negative**	-0.0088	0.0204	0.6669	-0.0487	0.0312
	Neutral	0.0332*	0.0134	0.0129	0.007	0.0594
	Positive	-0.0244	0.0203	0.2285	-0.0642	0.0153
Hate speech detection	**Hateful**	-0.0845***	0.0117	<0.001	-0.1074	-0.0615
	Targeted	-0.01**	0.0033	0.0027	-0.0165	-0.0034
	Aggressive	-0.0458***	0.0059	<0.001	-0.0572	-0.0343
Sentiment toward authority	**Authority/Subversion**	-0.0064**	0.0023	0.0054	-0.0109	-0.0019
	Authority	-0.0055*	0.0022	0.0108	-0.0097	-0.0013
	Subversion	-0.0073	0.0046	0.1097	-0.0163	0.0017

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests).

Discussion and Conclusion

This study uses individual-level Twitter data to investigate the socioeconomic inequality in social capital and communication behaviours. On the one hand, the paper studies

socioeconomic inequality in social capital on Twitter by testing the hypotheses on the relationship between SES and social capital across different measures of social capital. On the other hand, the paper explores the socioeconomic inequality in communication behaviour on Twitter through three angles: hypothesis testing, descriptive exploration of topics, and a case study.

Confirming all the hypotheses about the relationship between SES and social capital, the paper illustrates that high SES Twitter users tend to have higher social capital than low SES users across different social capital measures. Consistent with the findings of Eagle et al. (2010), Luo et al. (2017), and Norbutas and Corten (2018), the paper finds that high SES Twitter users tend to have higher degree centrality, topological diversity, and effective size than low SES users. The paper further extends the existing literature by showing that high SES Twitter users tend to have higher reciprocity, lower local clustering coefficient, and higher efficiency. These findings underscore the replicability of the positive relationship between SES and competitive social capital within Twitter's communication network, suggesting that socioeconomic inequality in social capital is mirrored both offline and online. The findings raise critical questions about the potential role of social capital in reinforcing socioeconomic disparities on social media, highlighting the need for further investigation into the underlying mechanisms.

The study also confirms all the hypotheses about the relationship between SES and communication patterns. It shows that high SES Twitter users tend to use more complex and future-oriented language in their tweets than low SES users. The results of language complexity are significant whether measured by the Flesch–Kincaid readability score or various measures of lexical diversity. Although the measures are simple, the findings are not trivial. All tweets in this study are constrained by the 280-character limit (before Twitter offered premium users longer text). The fact that these simple measures are significantly different by SES, even with such a short text window, reflects the pervasiveness of socioeconomic inequality in communication behaviours. Moreover, the existing differences in communication patterns may further reinforce socioeconomic inequality, as they are often used to assess people's SES and competency (Bradac et al. 1977; Kraus et al. 2017). Similar to the findings on social capital, the findings on communication behaviour indicate the role of communication behaviours in reinforcing socioeconomic inequality on social media platforms, encouraging deeper research in the area.

Next, the paper shows some descriptive differences between the topics tweeted by high SES and low SES users. The topics are identified by hashtags and topic modelling. Both high and low SES users use hashtags related to entertainment and Covid-19. Apart from those, high SES users mainly use hashtags related to politics, whereas low SES users mainly use hashtags about holidays (and promotions if we consider promotional tweets). There is no significant difference in the probability of tweeting about most topics identified by topic modelling. A notable topic is politics. While high SES users are more likely to use hashtags related to politics than low SES users, no significant differences are found in the probability of tweeting about politics-related topics generated from the topic modelling. This finding only partially confirmed existing literature that politics is more likely to be tweeted by high SES users (Abitbol and Morales 2021; Preoţiuc-Pietro, Volkova, et al. 2015). The inconsistency highlights the need for further research in this area and the importance of considering SES when conducting research on Twitter. Twitter users already tend to have higher SES than the general public (Wojcik and Hughes 2019). As this paper shows that hashtags related to politics are disproportionally used by high SES users, the generalisability of existing studies that used hashtags to select tweets to

study political discussions may be further constrained by using a sample that is even more skewed to high SES users than average Twitter users.

Lastly, this paper shows that, when tweeting about immigration, high SES Twitter users tend to be more neutral, less hateful, and less authoritarian than low SES users. This finding is particularly interesting given its contrast with the average trends in the sample, where high SES Twitter tend to be less positive, more negative, more hateful and more authoritarian. This contrast seems to correspond well with the findings from social psychology that while low SES people are more prosocial and empathetic than high SES people in general, they tend to hold more negative attitudes and authoritarian views about ethnic minorities and immigrants (Manstead 2018). However, the results of average positive-negative sentiments are inconsistent with Preoţiuc-Pietro, Volkova, et al.'s findings (2015), which show that high SES users tend to be more neutral, less positive, and less negative. The inconsistency may result from the biased sample in either or both studies, so further research is needed for generalisable and conclusive insights. Nevertheless, the findings demonstrate the usefulness of using data from social media platforms like Twitter to replicate or extend existing evidence on the relationship between SES and sentiments.

While these findings offer valuable insights, they need to be interpreted with several limitations in mind. First, the sample used in this study is limited due to the closure of Twitter Academic API. The study only includes a sample of users with the highest and lowest estimated SES. On the one hand, the sample may not be representative of Twitter users. On the other hand, although theoretically unlikely, the findings cannot rule out nonlinear relationships between SES and social capital or communication behaviour. Further research is needed with more representative data of Twitter users across the SES spectrum. While a few researchers may obtain access to representative proprietary data, more researchers could incorporate existing Twitter data archived by research institutions, utilise data donations, or conduct similar research with data from social media platforms with better API access, like Reddit.

Second, the potential of tautology needs to be further clarified. In this study, Twitter users' SES is estimated using the accounts they follow. Some may argue that the accounts people follow on Twitter may have more direction associations with their @mentions network and communication behaviours instead of through SES, making the findings tautological. To address this issue, this study uses matching to control for eight potential confounders that measure user activities. It is reasonable to assume that the confounders play a mediating role in the direct associations between the accounts people follow and their @mentions network or communication behaviour. Therefore, after controlling for those confounders, the remaining associations between the accounts people follow and their @mentions network or communication behaviour should be mainly driven by their associations with SES. It is possible that the accounts people follow may still have direct associations with the radius-1 social capital measures even after controlling for the confounders. However, the associations with radius-2 social capital, language complexity, future orientation, and sentiments toward immigration should be minimal. Therefore, most of the findings of this study are not affected by this potential of tautology. It is obvious that SES from external sources, such as linking Twitter with survey data, would be more convincing, but that requires resources that are not easily available. A contribution of this study is precisely to show the utility of SES in Twitter research, which hopefully justifies and encourages future research in linking Twitter with more established SES data like surveys at the individual level.

Despite the limitations, this study provides significant contributions to the literature. This paper shows that Twitter users with higher SES tend to have higher social capital and more advantageous communication behaviours. The findings encourage future research to examine the mechanisms of how social capital and communication behaviours reinforce socioeconomic inequality on Twitter or similar social media platforms. The paper also confirms the utility of the method for estimating the SES of Twitter users from He and Tsvetkova (2023) and highlights the importance of considering SES when researching Twitter or similar social media platforms. Future research could utilise the method to conduct deeper examinations of socioeconomic inequality or account for SES when studying other topics such as communication, politics, and health. The limitations of this study call for future research in linking data from social media platforms with traditional survey or administrative data at the individual level to get representative samples and more generalisable findings.

Acknowledgements

The author is grateful to Ana Macanovic for sharing data and code.

References

Abitbol, Jacob Levy, Eric Fleury, and Márton Karsai. 2019. 'Optimal Proxy Selection for Socioeconomic Status Inference on Twitter'. *Complexity* 2019(1):6059673. doi: 10.1155/2019/6059673.

Abitbol, Jacob Levy, and Alfredo J. Morales. 2021. 'Socioeconomic Patterns of Twitter User Activity'. *Entropy* 23(6):780. doi: 10.3390/e23060780.

Bernstein, Basil. 1960. 'Language and Social Class'. *The British Journal of Sociology* 11(3):271–76. doi: 10.2307/586750.

Bestgen, Yves. 2024. 'Measuring Lexical Diversity in Texts: The Twofold Length Problem'. *Language Learning* n/a(n/a). doi: 10.1111/lang.12630.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. 'Latent Dirichlet Allocation'. *Journal of Machine Learning Research* 3(4–5):993–1022.

Borgatti, Stephen P. 1997. 'Structural Holes: Unpacking Burt's Redundancy Measures'. *Connections* 20(1):35–38.

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

Bourdieu, Pierre. 1986. 'The Forms of Capital'. Pp. 241–58 in *Handbook of Theory and Research for the Sociology of Education*, edited by J. Richardson. New York: Greenwood.

Bradac, James J., Robert A. Davies, John A. Courtright, Roger J. Desmond, and Johnny I. Murdock. 1977. 'Richness of Vocabulary: An Attributional Analysis'. *Psychological Reports* 41(3_suppl):1131–34. doi: 10.2466/pr0.1977.41.3f.1131.

Burt, Ronald S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, Mass: Harvard University Press.

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. doi: 10.1038/s41586-022-04996-4.

Coleman, James S. 1988. 'Social Capital in the Creation of Human Capital'. *American Journal of Sociology* 94:S95–120. doi: 10.1086/228943.

DiMaggio, Paul, and Filiz Garip. 2012. 'Network Effects and Social Inequality'. *Annual Review of Sociology* 38(1):93–118. doi: 10.1146/annurev.soc.012809.102545.

Eagle, N., M. Macy, and R. Claxton. 2010. 'Network Diversity and Economic Development'. *Science* 328(5981):1029–31. doi: 10.1126/science.1186605.

Fagiolo, Giorgio. 2007. 'Clustering in Complex Directed Networks'. *Physical Review E* 76(2):026107. doi: 10.1103/PhysRevE.76.026107.

Flekova, Lucie, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. 'Exploring Stylistic Variation with Age and Income on Twitter'. Pp. 313–19 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics.

Gelderblom, Derik. 2018. 'The Limits to Bridging Social Capital: Power, Social Context and the Theory of Robert Putnam'. *The Sociological Review* 66(6):1309–24. doi: 10.1177/0038026118765360.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. 'Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism'. Pp. 55–130 in *Advances in Experimental Social Psychology*. Vol. 47, edited by P. Devine and A. Plant. Academic Press.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. 'Liberals and Conservatives Rely on Different Sets of Moral Foundations.' *Journal of Personality and Social Psychology* 96(5):1029–46. doi: 10.1037/a0015141.

Granovetter, Mark. 2005. 'The Impact of Social Structure on Economic Outcomes'. 18.

Granovetter, Mark S. 1973. 'The Strength of Weak Ties'. *American Journal of Sociology* 78(6):1360–80. doi: 10.1086/225469.

Hazen, Nancy L., and Betty Black. 1989. 'Preschool Peer Communication Skills: The Role of Social Status and Interaction Context'. *Child Development* 60(4):867–76. doi: 10.2307/1131028.

He, Yuanmo, and Milena Tsvetkova. 2023. ‘A Method for Estimating Individual Socioeconomic Status of Twitter Users’. *Sociological Methods & Research* 00491241231168665. doi: 10.1177/00491241231168665.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. ‘Matching as Nonparametric Pre-processing for Reducing Model Dependence in Parametric Causal Inference’. *Political Analysis* 15(3):199–236. doi: 10.1093/pan/mpl013.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. ‘MatchIt: Nonparametric Pre-processing for Parametric Causal Inference’. *Journal of Statistical Software* 42:1–28. doi: 10.18637/jss.v042.i08.

Hoover, Joe, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. ‘Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment’. *Social Psychological and Personality Science* 11(8):1057–71. doi: 10.1177/1948550619876629.

Ireland, Molly E., H. Andrew Schwartz, Qijia Chen, Lyle H. Ungar, and Dolores Albarracín. 2015. ‘Future-Oriented Tweets Predict Lower County-Level HIV Prevalence in the United States.’ *Health Psychology* 34(Suppl):1252–60. doi: 10.1037/hea0000279.

Jackson, Matthew O. 2021. ‘Inequality’s Economic and Social Roots: The Role of Social Networks and Homophily’.

Julien, Chris. 2015. ‘Bourdieu, Social Capital and Online Interaction’. *Sociology* 49(2):356–73. doi: 10.1177/0038038514535862.

Kincaid, J., Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. ‘Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel’. *Institute for Simulation and Training*.

Kraus, Michael W., Jun Won Park, and Jacinth J. X. Tan. 2017. ‘Signs of Social Class: The Experience of Economic Inequality in Everyday Life’. *Perspectives on Psychological Science* 12(3):422–35. doi: 10.1177/1745691616673192.

Lampos, Vasileios, Nikolaos Aletras, Jens K. Geyti, Bin Zou, and Ingemar J. Cox. 2016. ‘Inferring the Socioeconomic Status of Social Media Users Based on Behaviour and Language’. Pp. 689–95 in *Advances in Information Retrieval, Lecture Notes in Computer Science*, edited by N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello. Cham: Springer International Publishing.

Laureate, Caitlin Doogan Poet, Wray Buntine, and Henry Linger. 2023. ‘A Systematic Review of the Use of Topic Models for Short Text Social Media Analysis’. *Artificial Intelligence Review* 56(12):14223–55. doi: 10.1007/s10462-023-10471-x.

Lin, Nan. 1999. ‘Building a Network Theory of Social Capital’’. *Connections* 22(1):28–51.

Lin, Nan. 2000. ‘Inequality in Social Capital’. *Contemporary Sociology* 29(6):785–95. doi: 10.2307/2654086.

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. doi: 10.1038/ncomms15227.

Macanovic, Ana, and Wojtek Przepiorka. 2024. ‘A Systematic Evaluation of Text Mining Methods for Short Texts: Mapping Individuals’ Internal States from Online Posts’. *Behavior Research Methods*. doi: 10.3758/s13428-024-02381-9.

Malvern, David, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development*. London: Palgrave Macmillan UK.

Manstead, Antony S. R. 2018. ‘The Psychology of Social Class: How Socioeconomic Status Impacts Thought, Feelings, and Behaviour’. *British Journal of Social Psychology* 57(2):267–91. doi: https://doi.org/10.1111/bjso.12251.

Martini, Franziska, Paul Samula, Tobias R. Keller, and Ulrike Klinger. 2021. ‘Bot, or Not? Comparing Three Methods for Detecting Social Bots in Five Political Discourses’. *Big Data & Society* 8(2):20539517211033566. doi: 10.1177/20539517211033566.

McCarthy, Philip M., and Scott Jarvis. 2007. ‘Vocd: A Theoretical and Empirical Evaluation’. *Language Testing* 24(4):459–88. doi: 10.1177/0265532207080767.

McCarthy, Philip M., and Scott Jarvis. 2010. ‘MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment’. *Behavior Research Methods* 42(2):381–92. doi: 10.3758/BRM.42.2.381.

Menshikova, Anastasia, and Frank van Tubergen. 2022. ‘What Drives Anti-Immigrant Sentiments Online? A Novel Approach Using Twitter’. *European Sociological Review* 38(5):694–706. doi: 10.1093/esr/jcac006.

Mouw, Ted. 2003. ‘Social Capital and Finding a Job: Do Contacts Matter?’ *American Sociological Review* 68(6):868–98. doi: 10.1177/000312240306800604.

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. doi: 10.1016/j.socnet.2017.06.002.

Pérez, Juan Manuel, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2024. ‘Pysentimiento: A Python Toolkit for Opinion Mining and Social NLP Tasks’.

Portes, Alejandro. 1998. ‘Social Capital: Its Origins and Applications in Modern Sociology’. *Annual Review of Sociology* 24(1):1–24. doi: 10.1146/annurev.soc.24.1.1.

Preis, Tobias, Helen Susannah Moat, H. Eugene Stanley, and Steven R. Bishop. 2012. ‘Quantifying the Advantage of Looking Forward’. *Scientific Reports* 2(1):350. doi: 10.1038/srep00350.

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. doi: 10.1371/journal.pone.0138717.

Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Rowe, Francisco, Michael Mahony, Eduardo Graells-Garrido, Marzia Rango, and Niklas Sievers. 2021. ‘Using Twitter to Track Immigration Sentiment during Early Stages of the COVID-19 Pandemic’. *Data & Policy* 3:e36. doi: 10.1017/dap.2021.38.

Sankoff, David, and Réjean Lessard. 1975. ‘Vocabulary Richness: A Sociolinguistic Analysis’. *Science* 190(4215):689–90. doi: 10.1126/science.190.4215.689.

Sayyadiharikandeh, Mohsen, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. ‘Detection of Novel Social Bots by Ensembles of Specialized Classifiers’. Pp. 2725–32 in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*. New York, NY, USA: Association for Computing Machinery.

Watts, Duncan J., and Steven H. Strogatz. 1998. ‘Collective Dynamics of “Small-World” Networks’. *Nature* 393(6684):440–42. doi: 10.1038/30918.

Westlund, Hans, and Frane Adam. 2010. ‘Social Capital and Economic Performance: A Meta-Analysis of 65 Studies’. *European Planning Studies* 18(6):893–919. doi: 10.1080/09654311003701431.

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/).

Chapter 6 Conclusion

This PhD thesis uses digital trace data and computational methods to study socioeconomic inequality in daily behaviours and social interactions. It advances our understanding of how socioeconomic status (SES) is reflected and reinforced in consumption practices, social capital, and communication behaviours. Through the three empirical papers, the thesis contributes to the growing computational social science (CSS) literature by providing new methods, frameworks, and insights into studying socioeconomic inequality in the digital age.

Summary of key findings

The first empirical paper presents a method for estimating the individual SES of Twitter users. The paper builds on Bourdieu's (1984) theory of SES to argue that the commercial and entertainment accounts that Twitter users follow reflect their economic and cultural capital, which can be used to infer the users' SES. Compared to previous attempts, this method is theory-based, scalable, and requires minimal data and computational resources. The paper applies the method using 339 popular US brands to estimate the SES of almost 3.5 million Twitter users. The approach is validated against external datasets, including Facebook Marketing API, job titles listed in Twitter profiles, and a small survey sample. The results show that the proposed method effectively captures SES with significant correlations to income, occupation, and education, while being minimally correlated with other demographic variables. The findings support the underlying principle of the proposed method and justify further efforts to refine it at the individual level. The proposed method opens new opportunities for innovative social research on inequality on Twitter and similar online platforms.

The second paper reveals inequality in daily consumption using large-scale data of mobile-tracked visits. This study expands the theoretical and empirical understanding of consumption inequality by integrating two key consumption theories—cultural omnivorousness (Peterson 1992) and inconspicuous consumption (Berger and Ward 2010). The findings confirm that higher SES individuals exhibit more diverse consumption patterns, both in terms of the range of brands and the price levels of the stores they visit. The associations persist across different industries, although stronger in industries involving leisure and cultural expression than those concerning necessity goods. Notably, the observed patterns cannot be fully explained by geographic constraints, indicating deeper social and cultural factors at play. The findings illustrate and quantify socioeconomic divisions in daily consumption practices, bearing further evidence for the pervasiveness and inevitability of socioeconomic inequality in daily life. The findings also provide further empirical support to the methodological assumption that consumption preferences could serve as reliable indicators of SES.

The third paper applies the SES estimation method developed in the first paper to investigate socioeconomic inequality in social capital and communication behaviours on Twitter. The paper reveals that higher SES Twitter users have higher social capital across different measures of social capital and tend to use more complex and future-oriented language in their tweets. Also, while high and low SES users mostly talk about similar topics, they tend to use different hashtags and have divergent sentiments towards immigration. These results provide evidence that socioeconomic inequalities are not only reflected in social media interactions but may also be reinforced by them. The study highlights the need for further research to explore the

underlying mechanisms and integrate SES as a critical factor in social media studies. The findings also further establish the utility of the method proposed in the first paper.

Contributions

This PhD thesis makes several significant contributions to the research in socioeconomic inequality and computational social science. First, it introduces a new method for estimating individual SES on Twitter, an approach that has far-reaching implications for future research. It provides a theory-based, scalable, and replicable tool to infer SES from digital trace data. This contribution addresses a critical limitation in the field, where SES is often underrepresented in studies due to a lack of direct socioeconomic indicators in publicly available data.

Second, this thesis bridges traditional sociological theory with modern computational methods, grounding its empirical investigations in the theoretical frameworks of Bourdieu (1984) and others. Bourdieu's concept of economic, cultural, and social capital is a guiding framework throughout the thesis, allowing for a nuanced interpretation of how these forms of capital interact in the digital age. The first paper embeds firmly on Bourdieu. The second paper incorporates theories of conspicuous consumption, cultural capital, cultural omnivorousness, and inconspicuous consumption. The Third paper encompasses theories on competitive and cooperative social capital. The application of the theories provides nuanced interpretations of how these forms of capital interact in the digital age and offers analytical frameworks for future studies.

Third, the thesis introduces a direction of focus on consumption preferences and patterns in both measuring SES and studying inequality. The first paper leverages consumption preferences for SES estimation and confirms the close relationship between SES and consumption preferences in the digital space. The second paper further confirms and dissects such a relationship, highlighting the role of consumption preferences in reinforcing socioeconomic inequality.

Fourth, the empirical findings of this thesis underscore the role of digital platforms in reinforcing existing socioeconomic inequalities. Social media platforms like Twitter are not neutral spaces; they reflect and perpetuate offline socioeconomic disparities. This thesis demonstrates that SES influences not only who we interact with online but also how we communicate, consume, and even express opinions on social issues. These findings have important implications for both academic research and policymaking, suggesting that efforts to address inequalities must consider broader social and cultural factors.

Broader Implications

The results of this dissertation have important implications for understanding socioeconomic inequality in digital environments. As digital platforms play an increasingly central role in mediating social interactions, consumption, and communication, it is crucial to understand how they reflect and reinforce existing inequalities. This research highlights the need to consider SES as a critical factor in studies of digital behaviours. Many existing studies of social media behaviours overlook SES, which can lead to incomplete or misleading conclusions about the role of digital platforms in shaping social life. The method for estimating SES developed in this thesis provides a valuable tool for addressing this gap, enabling researchers to incorporate SES into analyses of digital trace data.

The dissertation also has broader implications for understanding the role of social media in public life. As platforms like Twitter continue to shape political discourse, social movements, and information dissemination, it is critical to understand how SES influences participation and engagement on these platforms. The findings of this thesis suggest that SES plays a significant role in shaping who participates in online discussions and how they participate. It has important implications for understanding the role of social media in democratic processes and public discourse, as it suggests that digital platforms may amplify the voices of higher-status individuals, further entrenching existing inequalities

From a policy perspective, the findings have important implications for addressing inequality. The evidence that higher SES individuals have higher social capital and engage in more advantageous communication behaviours online suggests that social media platforms may reinforce social stratification. Policymakers and platform designers should consider how platform algorithms, data access, and community features may contribute to the reinforcement of existing inequalities. The findings from both Twitter and mobile-tracking data suggest that policy interventions aimed at reducing inequality must address cultural and social factors alongside economic ones.

Limitations and Future Directions

While this thesis has made significant contributions to the field, several limitations should be acknowledged. First, the reliance on Twitter as the primary data source raises questions about the generalizability of the findings to other social media platforms. Twitter, while widely used in academic research, has a unique user base and communication style that may not be representative of other platforms like Facebook, Instagram, or TikTok. Future research should explore whether the patterns observed in this dissertation hold across different platforms and in different cultural contexts.

Second, the recent changes to Twitter's API present challenges and directions for future research. The discontinuation of the free Academic Track API, which provided large-scale access to Twitter data, limits the replicability of this study. While the findings remain valid and valuable, future research may need to explore alternative data sources or collaborate with platforms to gain access to the necessary data. The thesis also raises broader concerns about the accessibility of digital trace data for research. During this PhD project, the primary data source for the second paper, SafeGraph, also shifted from free academic access to a paid model. Although the pricing of SafeGraph data access is more reasonable than Twitter, these changes show the unreliability of propriety data and corporate goodwill. The academic community should find better ways to collaborate on collecting, curating, accessing, linking, and using digital trace data.

Third, while this dissertation shows how SES is associated with daily behaviours and social interactions, much is still to be learned about the underlying mechanisms of these relationships. For example, how do individual motivations, cultural norms, and platform algorithms interact to shape the observed patterns? Future research could employ experimental, causal or qualitative methods to explore these questions in more depth.

Finally, the landscape of computational social science and inequality is rapidly evolving. The emergence of generative AI has significant implications for society and research. The technology is already affecting the society in many ways, and it is set to have even broader

implications. For computational social scientists, it becomes harder to distinguish humans and machines. More efforts are required if researchers are exclusively interested in human behaviours, while more doors are opened for studying the dynamics of human-machine systems. Moreover, generative AI has a significant potential to drastically benefit those controlling relevant resources, leaving many people behind. Addressing the potential inequality arising from AI technologies represents an important direction for future research on socioeconomic inequality.

References

Berger, Jonah, and Morgan Ward. 2010. 'Subtle Signals of Inconspicuous Consumption'. *Journal of Consumer Research* 37(4):555–69. doi: 10.1086/655445.

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

Peterson, Richard A. 1992. 'Understanding Audience Segmentation: From Elite and Mass to Omnivore and Univore'. *Poetics* 21(4):243–58. doi: 10.1016/0304-422X(92)90008-Q.

Supplementary Material

Paper 1 (Chapter 3) Supplementary Material

Supplementary Table 1: Brands included in the study

Brand	Twitter account	Domain
7-Eleven	7eleven	supermarket
Kroger	kroger	supermarket
Target	Target	supermarket
Walmart	Walmart	supermarket
Albertsons	Albertsons	supermarket
Food Lion	FoodLion	supermarket
Sam's Club	SamsClub	supermarket
Amazon Fresh	AmazonFresh	supermarket
Aldi	AldiUSA	supermarket
Big Lots	BigLots	supermarket
Dollar General	DollarGeneral	supermarket
Dollar Tree	DollarTree	supermarket
Save-A-Lot	SaveALot	supermarket
Family Dollar	myfamilydollar	supermarket
central market	CentralMarket	supermarket
Whole Foods Market	WholeFoods	supermarket
Sprouts Farmers Market	sproutsfm	supermarket
Wegmans Food Markets	Wegmans	supermarket
Natural Grocers	NaturalGrocers	supermarket
Earth Fare	EarthFare	supermarket
Winn-Dixie	WinnDixie	supermarket
Publix	Publix	supermarket
WinCo Foods	WinCoFoods	supermarket
Giant Food Stores	GiantFoodStores	supermarket
Food City	FoodCity	supermarket
ShopRite	ShopRiteStores	supermarket
J. C. Penney	jcpenney	department store
Kohl's	Kohls	department store
Macy's	Macys	department store
Sears	Sears	department store
Bloomingdale's	Bloomingdales	department store
Neiman Marcus	neimanmarcus	department store
Bergdorf Goodman	Bergdorfs	department store
Nordstrom	Nordstrom	department store

Nordstrom Rack	nordstromrack	department store
Saks Fifth Avenue	saks	department store
Burlington	Burlington	department store
Five Below	fivebelow	department store
Fred Meyer	Fred_Meyer	department store
Gordmans	gordmans	department store
HomeGoods	HomeGoods	department store
Kmart	Kmart	department store
Marshalls	marshalls	department store
Meijer	meijer	department store
Ollie's Bargain Outlet	OlliesOutlet	department store
T.J. Maxx	tjmaxx	department store
Tuesday Morning	TuesdayMorning	department store
Belk	belk	department store
Century 21 Stores	century21stores	department store
Dillard's	Dillards	department store
Black & Decker	BLACKANDDECKER	speciality retail
Craftsman	craftsman	speciality retail
John Deere	JohnDeere	speciality retail
Sherwin-Williams	SherwinWilliams	speciality retail
DeWalt	DEWALTtough	speciality retail
Skil	SkilTools	speciality retail
Makita Tools	MakitaTools	speciality retail
RYOBI Tools	RYOBItoolsusa	speciality retail
Husqvarna	HusqvarnaUSA	speciality retail
RIDGID Tools	RIDGIDtoday	speciality retail
The Home Depot	HomeDepot	speciality retail
Lowe's	Lowes	speciality retail
Barnes & Noble	BNBuzz	speciality retail
Best Buy	BestBuy	speciality retail
Ace Hardware	AceHardware	speciality retail
Walgreens	Walgreens	speciality retail
Hallmark	Hallmark	speciality retail
Mattel	Mattel	speciality retail
La-Z-Boy	lazboy	speciality retail
StaplesStores	StaplesStores	speciality retail
CVS Pharmacy	cvspharmacy	speciality retail
Office Depot	officedepot	speciality retail
IKEA	IKEAUSA	speciality retail
PetSmart	PetSmart	speciality retail
Petco	Petco	speciality retail
Toys "R" Us	ToysRUs	speciality retail

True Value	TrueValue	speciality retail
Kohler	Kohler	speciality retail
houzz	houzz	speciality retail
Miele	MieleUSA	speciality retail
Yankee Candle	TheYankeeCandle	speciality retail
Sealy	Sealy	speciality retail
OfficeMax	OfficeMax	speciality retail
American Greetings	amgreetings	speciality retail
Tempur-Pedic	TempurPedic	speciality retail
Serta Mattress	SertaMattresses	speciality retail
Rite Aid	riteaid	speciality retail
RadioShack	RadioShack	speciality retail
Bass Pro Shops	BassProShops	speciality retail
DICK'S Sporting Goods	DICKS	speciality retail
Pier 1 Imports	pier1	speciality retail
Party City	PartyCity	speciality retail
Foot Locker	footlocker	speciality retail
Disney Store	shopDisney	speciality retail
Build-A-Bear Workshop	buildabear	speciality retail
Benjamin Moore	Benjamin_Moore	speciality retail
Cabela's	Cabelas	speciality retail
Tiffany & Co.	TiffanyAndCo	speciality retail
Pottery Barn	potterybarn	speciality retail
Sleep Number	sleepnumber	speciality retail
Moen	moen	speciality retail
GNC	GNCLiveWell	speciality retail
Carter's	Carters	speciality retail
BabiesRUs	BabiesRUs	speciality retail
Crate and Barrel	CrateandBarrel	speciality retail
Zalesjewelers	ZalesJewelers	speciality retail
Famous Footwear	FamousFootwear	speciality retail
Planet Fitness	PlanetFitness	speciality retail
Beautyrest	Beautyrest	speciality retail
Pandora jewelry	PANDORA_NA	speciality retail
Kay Jewelers	KayJewelers	speciality retail
Williams Sonoma	WilliamsSonoma	speciality retail
Ashley HomeStore	AshleyHomeStore	speciality retail
Claire's	claires	speciality retail
THe Children's Place	childrensplace	speciality retail
REI	REI	speciality retail
The Vitamin Shoppe	VitaminShoppe	speciality retail
American Girl	American_Girl	speciality retail

Champs Sports	champssports	speciality retail
Shoe Carnival	ShoeCarnival	speciality retail
Mattress Firm	MattressFirm	speciality retail
Jared	ThatsJared	speciality retail
Gymboree	Gymboree	speciality retail
Finish Line	FinishLine	speciality retail
LA Fitness	LAFitness	speciality retail
journeys	journeys	speciality retail
Academy Sports + Outdoors	Academy	speciality retail
Helzberg Diamonds	Helzberg	speciality retail
Abercrombie kids	abercrombiekids	speciality retail
Big 5 Sporting Goods	big5since55	speciality retail
Rack Room Shoes	myrackroomshoes	speciality retail
Modell's	Modells	speciality retail
Peloton	onepeloton	speciality retail
Havertys Furniture	havertys	speciality retail
Warby Parker	WarbyParker	speciality retail
Raymour & Flanigan	raymourflanigan	speciality retail
Zenni Optical	zennioptical	speciality retail
P. C. Richard & Son	PCRichardandSon	speciality retail
SoulCycle	soulcycle	speciality retail
Crunch	CrunchGym	speciality retail
Art Van Furniture	artvan	speciality retail
Kendra Scott	KendraScott	speciality retail
Flywheel Sports	Flywheel	speciality retail
Saatva	SaatvaMattress	speciality retail
Verizon Wireless	VZWSupport	speciality retail
AT&T Wireless	ATT	speciality retail
Build.com	buildcom	speciality retail
Harbor Freight Tools	HarborFreight	speciality retail
B&H Photo	BHPhotoVideo	speciality retail
Crutchfield	Crutchfield	speciality retail
Fry's Electronics	fryselectronics	speciality retail
Micro Center	microcenter	speciality retail
Newegg	Newegg	speciality retail
Bed Bath & Beyond	BedBathBeyond	speciality retail
BHG Live Better	BHGLiveBetter	speciality retail
Bath & Body Works	bathbodyworks	speciality retail
TigerDirect	TigerDirect	speciality retail
McDonald's	McDonalds	restaurant
Starbucks	Starbucks	restaurant
Subway	SUBWAY	restaurant

Taco Bell	tacobell	restaurant
Chick-fil-A	ChickfilA	restaurant
Burger King	BurgerKing	restaurant
Wendy's	Wendys	restaurant
Dunkin' Donuts	dunkindonuts	restaurant
Domino's	dominos	restaurant
Panera Bread	panerabread	restaurant
Pizza Hut	pizzahut	restaurant
Chipotle Mexican Grill	ChipotleTweets	restaurant
Sonic Drive-In	sonicdrivein	restaurant
KFC	kfc	restaurant
Applebee's	Applebees	restaurant
Olive Garden	olivegarden	restaurant
Arby's	Arbys	restaurant
Little Caesars	littlecaesars	restaurant
Buffalo Wild Wings	BWWings	restaurant
Dairy Queen	DairyQueen	restaurant
Panda Express	PandaExpress	restaurant
Maggiano's Little Italy	Maggianos	restaurant
Cold Stone Creamery	ColdStone	restaurant
Zoës Kitchen	ZoesKitchen	restaurant
PeiWei	PeiWei	restaurant
Caribou Coffee	cariboucoffee	restaurant
Philz Coffee	PhilzCoffee	restaurant
Peet's Coffee	peetscoffee	restaurant
Ben & Jerry's	benandjerrys	restaurant
Dippin' Dots	DippinDots	restaurant
Pinkberry	Pinkberry	restaurant
The Melting Pot	TheMeltingPot	restaurant
Brio Tuscan Grill	BrioItalian	restaurant
Tijuana Flats	TijuanaFlats	restaurant
Uno Pizzeria & Grill	UnoChicagoGrill	restaurant
Morton's The Steakhouse	Mortons	restaurant
The Cheesecake Factory	Cheesecake	restaurant
Red Mango	Red Mango	restaurant
P.F. Chang's China Bistro	PFChangs	restaurant
Jason's Deli	jasonsdeli	restaurant
Shake Shack	shakeshack	restaurant
The Capital Grille	CapitalGrille	restaurant
Souplantation	souplantation	restaurant
Buca di Beppo	bucadibeppo	restaurant
Jack in the Box	JackBox	restaurant

Denny's	DennysDiner	restaurant
LongHorn Steakhouse	LongHornSteaks	restaurant
Hooters	Hooters	restaurant
carl's jr	CarlsJr	restaurant
Nike	Nike	clothing
adidas	adidasUS	clothing
Levi's	LEVIS	clothing
Old Navy	OldNavy	clothing
Victoria's Secret	VictoriasSecret	clothing
Hanes	Hanes	clothing
Calvin Klein	CalvinKlein	clothing
Gap	Gap	clothing
Tommy Hilfiger	TommyHilfiger	clothing
SKECHERS	SKECHERSUSA	clothing
Converse	Converse	clothing
Ralph Lauren	RalphLauren	clothing
Wrangler Jeans	Wrangler	clothing
Banana Republic	BananaRepublic	clothing
Under Armour	UnderArmour	clothing
American Eagle	AEO	clothing
Lee Jeans	LeeJeans	clothing
Abercrombie & Fitch	Abercrombie	clothing
Crocs Shoes	Crocs	clothing
New Balance	newbalance	clothing
Champion	ChampionUSA	clothing
Ray-Ban	ray_ban	clothing
Dockers	Dockers	clothing
L.L.Bean	LLBean	clothing
Men's Wearhouse	menswearhouse	clothing
Eddie Bauer	eddiebauer	clothing
Timberland	Timberland	clothing
Urban Outfitters	UrbanOutfitters	clothing
Forever 21	Forever21	clothing
J.Crew	jcrew	clothing
The North Face	thenorthface	clothing
Hush Puppies	hushpuppies_usa	clothing
Samsonite	SamsoniteUSA	clothing
Aéropostale	Aeropostale	clothing
Lane Bryant	lanebryant	clothing
Vans	VANS_66	clothing
Nautica	nautica	clothing
Michael Kors	MichaelKors	clothing

Jos A Bank	JosABank	clothing
H&M	hmusa	clothing
Canada Goose	canadagoose	clothing
Anthropologie	Anthropologie	clothing
Free People	FreePeople	clothing
ASOS	ASOS_Us	clothing
Zara	zarausa	clothing
USA Today	USATODAY	newspapers
The Wall Street Journal	WSJ	newspapers
The New York Times	nytimes	newspapers
The Washington Post	washingtonpost	newspapers
The Hill	thehill	newspapers
National Enquirer	NatEnquirer	newspapers
Star Magazine	Star_News	newspapers
The Onion	TheOnion	newspapers
Fox News	FoxNews	news
MSNBC	MSNBC	news
CNN	CNN	news
NBC News	NBCNews	news
CBS News	CBSNews	news
NPR	NPR	news
ESPN	espn	news
Free Speech TV	freespeechtv	news
Fusion News	FusionNews	news
Newsmax	newsmax	news
world net daily	worldnetdaily	news
Daily Caller	DailyCaller	news
Natinal Football Leauge	NFL	sports
National Basketball Association	NBA	sports
Women's National Basketball Association	WNBA	sports
USA Basketball	usabasketball	sports
Major League Baseball	MLB	sports
NHL Hockey	NHL	sports
NASCAR	NASCAR	sports
PGA	ThePGA	sports
Major League Soccer	MLS	sports
United States men's national soccer team	USMNT	sports
United States women's national soccer team	USWNT	sports
NCAA Women’s Basketball	ncaawbb	sports
NCAA Men's Basketball	marchmadness	sports
NCAA Football	NCAAFootball	sports

America's Got Talent	AGT	tv shows
American Idol	AmericanIdol	tv shows
Keeping Up with the Kardashians	KUWTK	tv shows
Undercover Boss	undercover_cbs	tv shows
Little Big Shots	NBCLilBigShots	tv shows
Empire	EmpireFOX	tv shows
48 Hours	48hours	tv shows
Lethal Weapon	LethalWeaponFOX	tv shows
MacGyver	MacGyverCBS	tv shows
america's funniest home videos	AFVofficial	tv shows
Family Guy	FamilyGuyonFOX	tv shows
Hell's Kitchen	HellsKitchenFOX	tv shows
Hawaii Five-0	HawaiiFive0CBS	tv shows
Bob's Burgers	BobsBurgersFOX	tv shows
Code Black	CodeBlackCBS	tv shows
NCIS	NCIS_CBS	tv shows
NCIS: New Orleans	NCISNewOrleans	tv shows
Blue Bloods	BlueBloods_CBS	tv shows
Dr. Ken	DrKenABC	tv shows
Dancing with the Stars	DancingABC	tv shows
Criminal Minds	CrimMinds_CBS	tv shows
The Simpsons	TheSimpsons	tv shows
Modern Family	ModernFam	tv shows
New Girl	New_GirlTV	tv shows
black-ish	blackishabc	tv shows
The Goldbergs	TheGoldbergsABC	tv shows
Brooklyn Nine-Nine	nbcbrooklyn99	tv shows
Fresh off the Boat	FreshOffABC	tv shows
This Is Us	NBCThisisUs	tv shows
The Big Bang Theory	bigbangtheory	tv shows
Good Place	nbcthegoodplace	tv shows
The Middle	TheMiddle_ABC	tv shows
The Blacklist	NBCBlacklist	tv shows
Madam Secretary	MadamSecretary	tv shows
Blindspot	NBCBlindspot	tv shows
American Housewife	AmericanWifeABC	tv shows
Greys Anatomy	GreysABC	tv shows
Westworld	WestworldHBO	tv shows
Wheel of Fortune	WheelofFortune	tv shows
The Price Is Right	PriceIsRight	tv shows
Family Feud	FamilyFeud	tv shows
Law & Order	nbcsvu	tv shows

Mad Men	MadMen_AMC	tv shows
MythBusters	MythBusters	tv shows
Judge Judy	JudgeJudy	tv shows
Antiques Roadshow	RoadshowPBS	tv shows
Let's Make a Deal	letsmakeadeal	tv shows
The Bachelorette	BacheloretteABC	tv shows
The Bachelor	BachelorABC	tv shows
Burn Notice	Burn_NoticeTV	tv shows
The Good Wife	TheGoodWife_CBS	tv shows
Parks and Recreation	parksandrecnbc	tv shows
Parenthood	nbcparenthood	tv shows
Shark Tank	ABCSharkTank	tv shows
Pawn Stars	pawnstars	tv shows
60 Minutes	60Minutes	tv shows

Supplementary Table 2: Estimated SES of brands

Brand	Twitter Account	Estimated SES
SoulCycle	soulcycle	1.85026335
Flywheel Sports	Flywheel	1.84850849
Warby Parker	WarbyParker	1.77789634
Peloton	onepeloton	1.75546092
Philz Coffee	PhilzCoffee	1.71826598
The Hill	thehill	1.67910198
Mad Men	MadMen_AMC	1.67814998
NPR	NPR	1.64223592
Free Speech TV	freespeechtv	1.55785377
60 Minutes	60Minutes	1.50700276
United States women's national soccer team	USWNT	1.50175501
Fusion News	FusionNews	1.48134043
The Washington Post	washingtonpost	1.4431104
Parks and Recreation	parksandrecnbc	1.44261677
The Wall Street Journal	WSJ	1.42596708
Whole Foods Market	WholeFoods	1.41962995
The Onion	TheOnion	1.41867528
Good Place	nbcthegoodplace	1.40420906
MSNBC	MSNBC	1.40003074
Anthropologie	Anthropologie	1.38132738
The Good Wife	TheGoodWife_CBS	1.34587971
United States men's national soccer team	USMNT	1.34082223
New Girl	New_GirlTV	1.33670851
Parenthood	nbcparenthood	1.32648516

Bergdorf Goodman	Bergdorfs	1.32164187
The New York Times	nytimes	1.28699496
REI	REI	1.28531437
Madam Secretary	MadamSecretary	1.27126462
Free People	FreePeople	1.27076029
The Bachelorette	BacheloretteABC	1.24868015
The Bachelor	BachelorABC	1.24746711
NBC News	NBCNews	1.20703451
Westworld	WestworldHBO	1.205023
CBS News	CBSNews	1.19182388
Wegmans Food Markets	Wegmans	1.18815599
Major League Soccer	MLS	1.17580732
USA Today	USATODAY	1.172246
Michael Kors	MichaelKors	1.16965665
central market	CentralMarket	1.16369203
This Is Us	NBCThisisUs	1.14602723
CNN	CNN	1.14477287
Modern Family	ModernFam	1.14239562
NCAA Women's Basketball	ncaawbb	1.11891311
NCAA Men's Basketball	marchmadness	1.11586989
Daily Caller	DailyCaller	1.10816358
J.Crew	jcrew	1.09453048
Nordstrom Rack	nordstromrack	1.08926714
Women's National Basketball Association	WNBA	1.04589452
Nordstrom	Nordstrom	1.04015269
Tiffany & Co.	TiffanyAndCo	1.03180116
Canada Goose	canadagoose	1.03003394
Shark Tank	ABCSharkTank	1.01981892
Starbucks	Starbucks	0.98070193
MythBusters	MythBusters	0.97240767
black-ish	blackishabc	0.96140657
Greys Anatomy	GreysABC	0.96079491
Fox News	FoxNews	0.95663894
Amazon Fresh	AmazonFresh	0.95401671
Newsmax	newsmax	0.95305124
48 Hours	48hours	0.95161943
Brooklyn Nine-Nine	nbcbrooklyn99	0.94169089
NHL Hockey	NHL	0.94106513
Neiman Marcus	neimanmarcus	0.93597336
Zara	zarausa	0.93320216
The Capital Grille	CapitalGrille	0.92757961
ESPN	espn	0.92682636

Saks Fifth Avenue	saks	0.9151595
houzz	houzz	0.91230329
LA Fitness	LAFitness	0.91072335
Pinkberry	Pinkberry	0.87927058
NCAA Football	NCAAFootball	0.87052343
Shake Shack	shakeshack	0.86862352
National Basketball Association	NBA	0.86508046
Natinal Football Leauge	NFL	0.86192818
USA Basketball	usabasketball	0.86018859
Major League Baseball	MLB	0.85706323
National Enquirer	NatEnquirer	0.85065133
The North Face	thenorthface	0.84881224
Peet's Coffee	peetscoffee	0.84411372
PGA	ThePGA	0.827271
Law & Order	nbcsvu	0.82519541
Kendra Scott	KendraScott	0.82276068
Victoria's Secret	VictoriasSecret	0.81425224
Fresh off the Boat	FreshOffABC	0.81143109
Bloomingdale's	Bloomingdales	0.81013288
world net daily	worldnetdaily	0.80839197
Nike	Nike	0.80351438
Antiques Roadshow	RoadshowPBS	0.79526511
Ralph Lauren	RalphLauren	0.79186138
Natural Grocers	NaturalGrocers	0.78805646
The Big Bang Theory	bigbangtheory	0.7862219
The Blacklist	NBCBlacklist	0.77303904
Dancing with the Stars	DancingABC	0.76411479
B&H Photo	BHPhotoVideo	0.76385452
Caribou Coffee	cariboucoffee	0.76106756
Keeping Up with the Kardashians	KUWTK	0.74573994
Zoes Kitchen	ZoesKitchen	0.74475319
The Middle	TheMiddle_ABC	0.74061767
Criminal Minds	CrimMinds_CBS	0.72746547
Blue Bloods	BlueBloods_CBS	0.72689642
Calvin Klein	CalvinKlein	0.70210159
Crunch	CrunchGym	0.6963711
Hawaii Five-0	HawaiiFive0CBS	0.68586622
Blindspot	NBCBlindspot	0.67648299
NCIS	NCIS_CBS	0.66303445
American Idol	AmericanIdol	0.64923378
Empire	EmpireFOX	0.64373267
The Goldbergs	TheGoldbergsABC	0.64142855

Tijuana Flats	TijuanaFlats	0.63433134
Publix	Publix	0.63003714
Burn Notice	Burn_NoticeTV	0.59944289
Urban Outfitters	UrbanOutfitters	0.59934359
Tommy Hilfiger	TommyHilfiger	0.59911987
Under Armour	UnderArmour	0.58255576
Code Black	CodeBlackCBS	0.57607913
Ray-Ban	ray_ban	0.56613719
Ben & Jerry's	benandjerrys	0.56204359
IKEA	IKEAUSA	0.55627358
Forever 21	Forever21	0.55175518
Target	Target	0.54326341
Dr. Ken	DrKenABC	0.53506866
America's Got Talent	AGT	0.50564868
Bob's Burgers	BobsBurgersFOX	0.50209274
Judge Judy	JudgeJudy	0.48439277
Banana Republic	BananaRepublic	0.47862826
Family Guy	FamilyGuyonFOX	0.47066041
Verizon Wireless	VZWSupport	0.45817429
Century 21 Stores	century21stores	0.44786672
adidas	adidasUS	0.44567847
Morton's The Steakhouse	Mortons	0.42871717
Crate and Barrel	CrateandBarrel	0.41738522
American Housewife	AmericanWifeABC	0.39975337
H&M	hmusa	0.37967957
Gap	Gap	0.36983575
Macy's	Macys	0.3636703
The Simpsons	TheSimpsons	0.36313679
New Balance	newbalance	0.36064484
Barnes & Noble	BNBuzz	0.35647192
NCIS: New Orleans	NCISNewOrleans	0.35097975
NASCAR	NASCAR	0.3268435
AT&T Wireless	ATT	0.32034972
Hell's Kitchen	HellsKitchenFOX	0.30765162
Chipotle Mexican Grill	ChipotleTweets	0.30472548
Vans	VANS_66	0.28787354
L.L.Bean	LLBean	0.2813081
Sprouts Farmers Market	sproutsfm	0.25885385
Undercover Boss	undercover_cbs	0.25086291
Williams Sonoma	WilliamsSonoma	0.24725512
GNC	GNCLiveWell	0.2437662
Converse	Converse	0.23648098

Levi's	LEVIS	0.22437802
MacGyver	MacGyverCBS	0.2210695
Pawn Stars	pawnstars	0.17808774
HomeGoods	HomeGoods	0.1746691
Maggiano's Little Italy	Maggianos	0.16805615
Foot Locker	footlocker	0.16333404
Planet Fitness	PlanetFitness	0.15617825
Benjamin Moore	Benjamin_Moore	0.14965864
Giant Food Stores	GiantFoodStores	0.14446553
T.J. Maxx	tjmaxx	0.11898885
Lethal Weapon	LethalWeaponFOX	0.11891277
Miele	MieleUSA	0.11731011
Star Magazine	Star_News	0.09963944
Little Big Shots	NBCLilBigShots	0.08619006
Disney Store	shopDisney	0.07974276
John Deere	JohnDeere	0.07281257
Chick-fil-A	ChickfilA	0.07092739
Wendy's	Wendys	0.06571195
DICK'S Sporting Goods	DICKS	0.06474871
Subway	SUBWAY	0.03576481
Nautica	nautica	0.03281331
Pottery Barn	potterybarn	0.0306995
Timberland	Timberland	0.03040127
Dunkin' Donuts	dunkindonuts	0.01766488
McDonald's	McDonalds	0.01576095
Finish Line	FinishLine	0.01313203
Walgreens	Walgreens	0.01085511
Food City	FoodCity	0.00250669
Best Buy	BestBuy	0.00139298
Souplantation	souplantation	0.00103476
Zenni Optical	zennioptical	-0.0018818
Walmart	Walmart	-0.0070676
Abercrombie kids	abercrombiekids	-0.0074558
america's funniest home videos	AFVofficial	-0.0145266
Marshalls	marshalls	-0.0302953
Lane Bryant	lanebryant	-0.0330964
Old Navy	OldNavy	-0.0331065
Uno Pizzeria & Grill	UnoChicagoGrill	-0.0448954
Family Feud	FamilyFeud	-0.0553707
Buca di Beppo	bucadibeppo	-0.0602764
Buffalo Wild Wings	BWWings	-0.0645261
Toys "R" Us	ToysRUs	-0.0711652

Kohl's	Kohls	-0.0800244
7-Eleven	7eleven	-0.0800863
Champs Sports	champssports	-0.0908063
Fred Meyer	Fred_Meyer	-0.0966728
WinCo Foods	WinCoFoods	-0.1019931
Pier 1 Imports	pier1	-0.1064464
Panera Bread	panerabread	-0.1070209
The Price Is Right	PriceIsRight	-0.1099298
Let's Make a Deal	letsmakeadeal	-0.1112184
Academy Sports + Outdoors	Academy	-0.115978
PetSmart	PetSmart	-0.1178012
Eddie Bauer	eddiebauer	-0.1415246
American Eagle	AEO	-0.1475302
Dillard's	Dillards	-0.1489629
Taco Bell	tacobell	-0.1572675
Kroger	kroger	-0.1651818
ShopRite	ShopRiteStores	-0.1861987
Abercrombie & Fitch	Abercrombie	-0.1988999
Hooters	Hooters	-0.2070759
Earth Fare	EarthFare	-0.2160198
The Home Depot	HomeDepot	-0.2207653
J. C. Penney	jcpenney	-0.2239756
Kohler	Kohler	-0.2508241
Petco	Petco	-0.2555438
Jason's Deli	jasonsdeli	-0.2577426
Cabela's	Cabelas	-0.2871946
Bath & Body Works	bathbodyworks	-0.2872328
PeiWei	PeiWei	-0.3045129
Burger King	BurgerKing	-0.3064976
CVS Pharmacy	cvspharmacy	-0.3142802
ASOS	ASOS_Us	-0.3165329
Aldi	AldiUSA	-0.3426959
Sherwin-Williams	SherwinWilliams	-0.3458349
Albertsons	Albertsons	-0.3562105
Meijer	meijer	-0.3760322
Food Lion	FoodLion	-0.4077741
Lowe's	Lowes	-0.4093329
Wheel of Fortune	WheelofFortune	-0.4104991
Pizza Hut	pizzahut	-0.4227817
The Cheesecake Factory	Cheesecake	-0.4288108
Gordmans	gordmans	-0.4307637
Domino's	dominos	-0.4326207

Arby's	Arbys	-0.4384803
KFC	kfc	-0.4388844
Harbor Freight Tools	HarborFreight	-0.4430725
StaplesStores	StaplesStores	-0.4731852
Men's Wearhouse	menswearhouse	-0.4912771
Big Lots	BigLots	-0.5008094
Havertys Furniture	havertys	-0.5043303
RadioShack	RadioShack	-0.5079513
Belk	belk	-0.5130448
Winn-Dixie	WinnDixie	-0.5147324
Sam's Club	SamsClub	-0.5166391
Bass Pro Shops	BassProShops	-0.522056
The Melting Pot	TheMeltingPot	-0.5285621
The Vitamin Shoppe	VitaminShoppe	-0.5287133
Mattel	Mattel	-0.5374085
Moen	moen	-0.5618296
Mattress Firm	MattressFirm	-0.5679229
Dollar Tree	DollarTree	-0.5704792
P.F. Chang's China Bistro	PFChangs	-0.5791364
Ace Hardware	AceHardware	-0.5895613
Aeropostale	Aeropostale	-0.6095851
Dairy Queen	DairyQueen	-0.6101505
Olive Garden	olivegarden	-0.6131105
Art Van Furniture	artvan	-0.6132162
Sonic Drive-In	sonicdrivein	-0.6428079
Jack in the Box	JackBox	-0.6700965
Applebee's	Applebees	-0.6734553
Denny's	DennysDiner	-0.6772794
Ashley HomeStore	AshleyHomeStore	-0.6871724
Sears	Sears	-0.6977301
Dippin' Dots	DippinDots	-0.7074017
Burlington	Burlington	-0.7192496
LongHorn Steakhouse	LongHornSteaks	-0.7201628
Pandora jewelry	PANDORA_NA	-0.798734
carl's jr	CarlsJr	-0.8049766
BabiesRUs	BabiesRUs	-0.8082409
Champion	ChampionUSA	-0.8142758
Bed Bath & Beyond	BedBathBeyond	-0.8165318
Office Depot	officedepot	-0.8260922
Build-A-Bear Workshop	buildabear	-0.8735539
THe Children's Place	childrensplace	-0.8851875
Kmart	Kmart	-0.8886847

Husqvarna	HusqvarnaUSA	-0.8913174
Little Caesars	littlecaesars	-0.9007366
Rite Aid	riteaid	-0.9109715
Crocs Shoes	Crocs	-0.9224484
Panda Express	PandaExpress	-0.9486957
Tempur-Pedic	TempurPedic	-0.9518174
American Girl	American_Girl	-0.9689147
OfficeMax	OfficeMax	-0.9728119
Hallmark	Hallmark	-1.0032182
Zalesjewelers	ZalesJewelers	-1.0108272
Jos A Bank	JosABank	-1.0383709
journeys	journeys	-1.0453606
Dollar General	DollarGeneral	-1.0454884
Ollie's Bargain Outlet	OlliesOutlet	-1.0488593
Modell's	Modells	-1.0904362
Cold Stone Creamery	ColdStone	-1.1029376
True Value	TrueValue	-1.1040892
Famous Footwear	FamousFootwear	-1.1138426
Lee Jeans	LeeJeans	-1.1414739
Yankee Candle	TheYankeeCandle	-1.1695267
Five Below	fivebelow	-1.17045
Beautyrest	Beautyrest	-1.2033161
Kay Jewelers	KayJewelers	-1.2110389
Gymboree	Gymboree	-1.2318941
Newegg	Newegg	-1.253648
Sealy	Sealy	-1.274264
Sleep Number	sleepnumber	-1.3146043
Rack Room Shoes	myrackroomshoes	-1.3348014
Party City	PartyCity	-1.4390457
Carter's	Carters	-1.4913769
Family Dollar	myfamilydollar	-1.5066386
Shoe Carnival	ShoeCarnival	-1.5237141
Makita Tools	MakitaTools	-1.5427004
RIDGID Tools	RIDGIDtoday	-1.5750707
TigerDirect	TigerDirect	-1.692482
Claire's	claires	-1.7041191
Jared	ThatsJared	-1.707567
Hanes	Hanes	-1.7150215
American Greetings	amgreetings	-1.7357701
Skil	SkilTools	-1.73863
Samsonite	SamsoniteUSA	-1.7428151
DeWalt	DEWALTtough	-1.7584897

Craftsman	craftsman	-1.7673461
Raymour & Flanigan	raymourflanigan	-1.8391125
Brio Tuscan Grill	BrioItalian	-1.8517373
Big 5 Sporting Goods	big5since55	-1.8635366
Dockers	Dockers	-1.8875077
SKECHERS	SKECHERSUSA	-1.8929584
RYOBI Tools	RYOBItoolsusa	-1.9111581
Build.com	buildcom	-2.0020564
La-Z-Boy	lazboy	-2.1256093
Helzberg Diamonds	Helzberg	-2.15038
Crutchfield	Crutchfield	-2.1797223
Black & Decker	BLACKANDDECKER	-2.2467993
Save-A-Lot	SaveALot	-2.3030819
BHG Live Better	BHGLiveBetter	-2.3079619
Wrangler Jeans	Wrangler	-2.3177918
Micro Center	microcenter	-2.4252766
Fry's Electronics	fryselectronics	-2.4284073
Serta Mattress	SertaMattresses	-2.5812379
P. C. Richard & Son	PCRichardandSon	-2.5817136
Tuesday Morning	TuesdayMorning	-2.6479654
Hush Puppies	hushpuppies_usa	-2.9543234

Supplementary Table 3: Selected job titles

Job title	Salary	EGP class
managing director	193,850	1
general manager	123,030	1
operations manager	123,030	1
sales manager	141,690	1
surgeon	252,040	2
dentist	183,060	2
lawyer	145,300	2
graphic designer	56,510	2
software developer	106,980	2
software engineer	106,980	2
registered nurse	77,460	2
marketing manager	149,200	2
art director	109,600	2
civil engineer	94,360	2
data scientist	100,560	2
physical therapist	90,170	2

nurse practitioner	111,840	2
paralegal	55,020	2
personal trainer	45,110	3
marketing specialist	71,570	3
estate agent	66,100	3
fashion designer	86,110	3
interior designer	60,990	3
event planner	54,880	3
massage therapist	47,180	3
firefighter	54,650	3
police officer	67,620	3
receptionist	31,250	4
locksmith	44,460	5
chef	56,310	5
carpenter	52,850	5
woodworker	34,660	5
tailor	33,950	5
florist	29,760	5
plumber	59,800	5
farmer	80,360	5
hairstylist	31,710	6
flight attendant	56,230	6
barber	31,710	6
travel agent	44,690	6
dog trainer	36,240	6
lifeguard	25,380	6
janitor	30,010	6
dispatcher	44,170	7
insurance agent	67,780	7
truck driver	42,170	8
bus driver	45,820	8
security guard	33,030	9
bartender	28,000	9
waiter	26,800	9
waitress	26,800	9
barista	25,020	9

Supplementary Table 4: Unavailable brands from Facebook Marketing API in the Divergent section

Brand	Domain	Estimated SES
Hanes	clothing	-1.7150215
Dockers	clothing	-1.8875077
Century 21 Stores	department store	0.44786672
Gordmans	department store	-0.4307637
Free Speech TV	news	1.55785377
Fusion News	news	1.48134043
MSNBC	news	1.40003074
NBC News	news	1.20703451
CBS News	news	1.19182388
CNN	news	1.14477287
Daily Caller	news	1.10816358
Fox News	news	0.95663894
Newsmax	news	0.95305124
world net daily	news	0.80839197
The Hill	newspapers	1.67910198
The Washington Post	newspapers	1.4431104
The Wall Street Journal	newspapers	1.42596708
The Onion	newspapers	1.41867528
The New York Times	newspapers	1.28699496
USA Today	newspapers	1.172246
Zoes Kitchen	restaurant	0.74475319
Souplantation	restaurant	0.00103476
Brio Tuscan Grill	restaurant	-1.8517373
Flywheel Sports	speciality retail	1.84850849
Warby Parker	speciality retail	1.77789634
Zenni Optical	speciality retail	-0.0018818
Toys "R" Us	speciality retail	-0.0711652
Pier 1 Imports	speciality retail	-0.1064464
Art Van Furniture	speciality retail	-0.6132162
BabiesRUs	speciality retail	-0.8082409
Makita Tools	speciality retail	-1.5427004
Jared	speciality retail	-1.707567
BHG Live Better	speciality retail	-2.3079619
NCAA Women's Basketball	sports	1.11891311

Natinal Football Leauge	sports	0.86192818
central market	supermarket	1.16369203
60 Minutes	tv shows	1.50700276
Parenthood	tv shows	1.32648516
48 Hours	tv shows	0.95161943
Blindspot	tv shows	0.67648299
Burn Notice	tv shows	0.59944289
Code Black	tv shows	0.57607913
American Housewife	tv shows	0.39975337
NCIS: New Orleans	tv shows	0.35097975

Note: Among these brands, *BHG Live Better, Zoes Kitchen, Souplantation, Code Black, NCIS: New Orleans, Blindspot, Burn Notice, Parenthood, BrioItalian, MakitaTools* returns audience size that are universally 1000 in multiple categories. The others do not return any results from the API.

Supplementary Table 5: The associations between education/income and other demographic variables in the YouGov survey data

		education			**income**		
	Test	**value**	**p**	**n**	**value**	**p**	**n**
Age	Spearman's correlation(ρ)	-0.057	0.466	162	0.016	0.848	147
Gender	t-test(t)	1.090	0.276	162	-0.349	0.728	147
Political Ideology	Spearman's correlation(ρ)	0.031	0.697	157	0.034	0.685	143
Race	Analysis of variance (F)	0.444	0.722	162	0.09	0.965	147

Supplementary Table 6: Results of regression analyses that predict the estimated SES with education/income, controlling for other demographic variables in the YouGov survey data.

	Model 1	Model 2
Education	0.385***	
	(0.107)	
Income		0.122*
		(0.051)
Age	0.016	0.008
	(0.011)	(0.011)
Gender (ref: Male)		
Female	-0.157	-0.238
	(0.311)	(0.330)
Race (ref: While)		
Black	0.293	0.340
	(0.588)	(0.600)
Hispanic	-0.484	-0.449
	(0.636)	(0.771)
Asian/other	0.131	0.274
	(0.751)	(0.905)
Political Ideology	-0.113	-0.077
	(0.115)	(0.126)
Intercept	-0.935	0.278
	(0.835)	(0.821)
n	157	143

Note: *p < 0.05, **p < 0.01 ***p < 0.001

Paper 2 (Chapter 4) Appendices

Appendix A. Additional information about the data sample.

Table A1. Four-digit NAICS industries and number of POIs included in the study.

Four-digit industry name	NAIC code	Number of POIs
Restaurants and Other Eating Places	7225	87070
Personal Care Services	8121	38452
Other Amusement and Recreation Industries	7139	15312
Grocery Stores	4451	15219
Health and Personal Care Stores	4461	13123
Clothing Stores	4481	12755
Gasoline Stations	4471	10356
Museums, Historical Sites, and Similar Institutions	7121	8360
Drinking Places (Alcoholic Beverages)	7224	6964
Sporting Goods, Hobby, and Musical Instrument Stores	4511	6935
Jewelry, Luggage, and Leather Goods Stores	4483	6906
Other Miscellaneous Store Retailers	4539	6901
Specialty Food Stores	4452	6430
Beer, Wine, and Liquor Stores	4453	4487
Florists	4531	4118
Used Merchandise Stores	4533	2972
Office Supplies, Stationery, and Gift Stores	4532	2949
Consumer Goods Rental	5322	2784
Shoe Stores	4482	2682
General Merchandise Stores, including Warehouse Clubs and Supercenters	4523	2503
Book Stores and News Dealers	4512	1946
Special Food Services	7223	1720
Amusement Parks and Arcades	7131	1272
Department Stores	4522	932
Motion Picture and Video Industries	5121	709
Gambling Industries	7132	491
Performing Arts Companies	7111	288
Spectator Sports	7112	190

As shown in Figure A1, the selected 13,653 CBGs slightly oversample lower income CBGs but still cover the full range of income levels. As our results show positive correlation between income and diversity, there is no reason to believe that the slight oversampling of lower income CBGs could significantly bias our results.

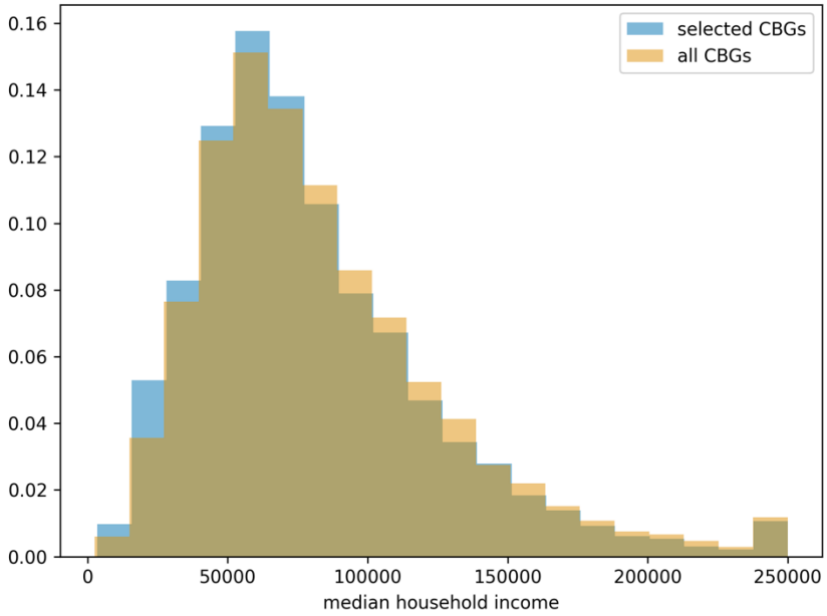

Figure A1. The distribution of median household income of all 52,509 CBGs in our sample versus the selected 13,653 CBGs that have at least 100 outgoing visitors.

We tried six measures of diversity, including the number of brands visited (*N brands* in Table A2), the normalized Shannon entropy by brands (*Brand entropy*), the range of brands' SES (*SES range*), the standard deviation of brands' SES (*SES std*), the number of price levels visited (*N price level*), and normalized Shannon entropy by price levels (*Price entropy*). As Table A2 shows, all measures are significantly correlated with each other and the median income of the CBG. Among the most correlated pairs (N brands & Brand entropy, SES range & SES std, and N price level & Price entropy), we chose the ones that represent more information to report and use in further analyses.

Table A2. Correlations between median household income and six different measures of diversity in consumption.

	Income	**N brands**	**Brand entropy**	**SES range**	**SES std**	**N price level**	**Price entropy**
Income	1	0.288***	0.292***	0.365***	0.471***	0.296***	0.358***
N brands	0.288***	1	0.918***	0.507***	0.321***	0.459***	0.262***
Brand entropy	0.292***	0.918***	1	0.543***	0.398***	0.493***	0.369***
SES range	0.365***	0.507***	0.543***	1	0.809***	0.398***	0.409***
SES std	0.471***	0.321***	0.398***	0.809***	1	0.363***	0.466***
N price level	0.296***	0.459***	0.493***	0.398***	0.363***	1	0.742***
Price entropy	0.358***	0.262***	0.369***	0.409***	0.466***	0.742***	1

Note: *** $p < 0.001$ (two-tailed tests).

Appendix B. Predicting CBGs' median household income from consumption patterns.

We use a LASSO regression model to predict the median household income of the CBGs with the proportion of outgoing visitors of the CBGs to the 924 brands. Following the standard practice, we keep 20 percent of the observations as a test set to examine the out-of-sample prediction accuracy. The LASSO regression model is an extension of the linear regression model estimated with ordinary least squares (OLS) that reduces overfitting and variance and increases prediction accuracy and interpretability. It adds a penalty term to the loss function of OLS regression which shrinks some of the coefficients towards zero. Let us take as an example the multiple linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \mathcal{E}$$

where y represents the outcome variable, β_0 is the intercept, x_j represents the *j*th predictor with coefficient β_j, $\mathcal{E}$ represents the error term, and p – the number of predictors. The OLS regression model aims to minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

where y_i represents the actual value of the *i*th observation of y, $\hat{y}_i$ represents the predicted value, and *n* represents the number of observations. The LASSO model, however, aims to minimize the following function:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is a tuning parameter to be determined separately. In this paper, we use the standard practice of using five-fold cross validation to select the value of λ. Due to the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$, to minimise the loss function, some coefficients need to shrink, sometimes all the way to zero. The shrinkage prevents overfitting and performs variable selection.

The LASSO regression model that predicts CBGs' median household income based on the proportion of outgoing visitors of the CBGs to the 924 brands has strong predictive power. For the training sample, the model explains 0.590 of the variances. Using the 355 brands selected by the model, the out-of-sample correlation between the predicted income and actual income is 0.748 ($p < 0.001$). Figure B1 shows the correlation between predicted and actual income in the test set. Such strong prediction performance suggests that SES is strongly associated with consumption preferences, even if consumption preferences are not determined by the economic constraints driven by the product prices. Nevertheless, the figure also shows that the predictions perform notably worse for CBGs with the highest income. One possible explanation for this is that the wealthiest individuals seek elite and authentic consumption experience and thus avoid brands.

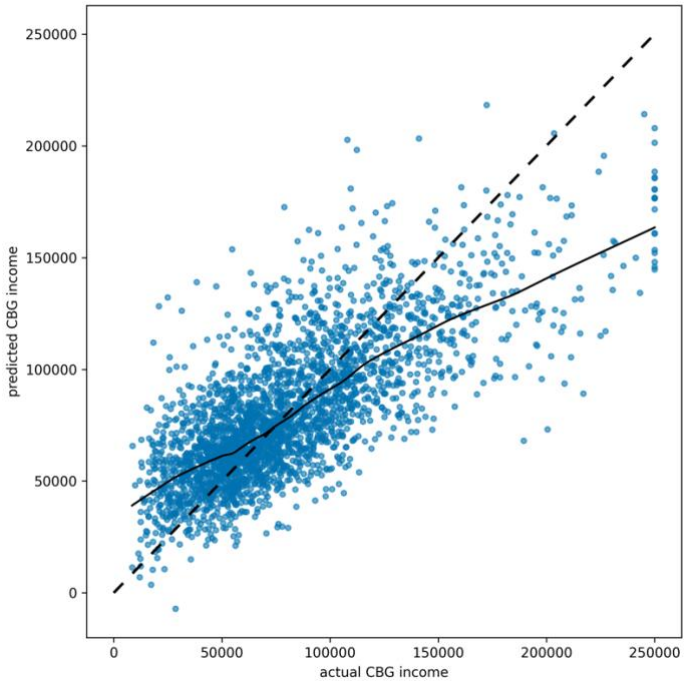

Figure B1. Correlation between predicted and actual income in the test set. The solid line shows the LOESS fit and the dashed line – the ideal 1:1 relation.

Appendix C. Results from separate regression analyses by industry that predict CBGs' diversity in consumption with median household income, controlling for income variability, estimated mobility, local availability, and demographic variables.

Table C1. Entropy by brand

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.406***	0.376***	0.440***	0.347***	0.437***	0.340***	0.260***	0.408***	0.289***
	(0.045)	(0.060)	(0.025)	(0.035)	(0.026)	(0.041)	(0.040)	(0.055)	(0.077)
Income variability	-0.120***	-0.136**	-0.115***	-0.106***	-0.080***	-0.051	-0.090**	-0.146***	-0.254***
	(0.034)	(0.042)	(0.019)	(0.025)	(0.021)	(0.034)	(0.029)	(0.039)	(0.061)
Mobility	0.013	-0.074	-0.106***	0.040	0.123***	-0.027	0.171***	0.101	0.012
	(0.052)	(0.043)	(0.020)	(0.022)	(0.024)	(0.027)	(0.034)	(0.058)	(0.069)
Local availability	0.135***	0.205***	0.186***	0.028	0.120***	0.073***	0.152***	0.122***	0.127*
	(0.031)	(0.041)	(0.015)	(0.020)	(0.015)	(0.020)	(0.024)	(0.036)	(0.049)
In NYC	0.224**	0.437***	0.118**	-0.159**	-0.387***	-0.508***	0.329***	-0.597***	-0.204
	(0.079)	(0.108)	(0.042)	(0.057)	(0.049)	(0.071)	(0.064)	(0.099)	(0.132)
Median age	0.015	0.066	0.005	0.004	0.002	0.033	-0.045	0.023	0.157**
	(0.031)	(0.042)	(0.016)	(0.021)	(0.017)	(0.021)	(0.025)	(0.037)	(0.048)
Proportion of male	-0.005	0.022	-0.020	-0.026	-0.030	-0.034	-0.046	0.014	0.083
	(0.029)	(0.040)	(0.015)	(0.020)	(0.016)	(0.020)	(0.025)	(0.035)	(0.045)
Proportion of white	-0.087	-0.140*	-0.174***	-0.150***	-0.126***	-0.093**	-0.034	-0.067	-0.178*
	(0.046)	(0.063)	(0.021)	(0.028)	(0.022)	(0.031)	(0.034)	(0.052)	(0.074)
Proportion of bachelor's degree or higher	0.080	-0.002	-0.092***	-0.157***	-0.247***	-0.077*	-0.076*	-0.057	0.097
	(0.042)	(0.060)	(0.022)	(0.031)	(0.025)	(0.034)	(0.035)	(0.052)	(0.069)
Intercept	0.163***	0.075	0.083***	0.150***	0.344***	0.282***	-0.013	0.500***	0.373***
	(0.042)	(0.063)	(0.021)	(0.028)	(0.019)	(0.029)	(0.033)	(0.045)	(0.065)
Observations	814	523	3192	2071	2559	1217	1578	608	362
R^2	0.264	0.215	0.191	0.058	0.162	0.123	0.097	0.187	0.138
Adjusted R^2	0.256	0.202	0.188	0.054	0.159	0.116	0.092	0.175	0.116
Residual Std. Error	0.828	0.926	0.857	0.894	0.763	0.698	0.943	0.858	0.916
F Statistic	32.075***	15.656***	83.301***	14.218***	54.602***	18.749***	18.742***	15.279***	6.261***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Table C2. Standard deviation of brands' SES

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.078	0.190*	0.272***	0.276***	0.006	0.218**	0.275***	0.231*	0.429**
	(0.076)	(0.081)	(0.032)	(0.067)	(0.054)	(0.072)	(0.075)	(0.096)	(0.127)
Income variability	0.069	0.007	0.060*	-0.032	-0.061	0.112	-0.089	0.152*	-0.102
	(0.061)	(0.057)	(0.025)	(0.048)	(0.044)	(0.061)	(0.055)	(0.068)	(0.083)
Mobility	0.272*	-0.050	-0.032	0.092	0.034	0.030	0.211**	0.053	0.388**
	(0.130)	(0.069)	(0.027)	(0.065)	(0.056)	(0.059)	(0.076)	(0.125)	(0.123)
Local availability	0.145*	0.131*	0.288***	0.264***	0.366***	0.091*	0.275***	0.176**	0.357***
	(0.058)	(0.058)	(0.021)	(0.040)	(0.031)	(0.039)	(0.048)	(0.064)	(0.076)
In NYC	0.141	0.154	-0.389***	-0.223	-0.293**	-0.030	0.050	-0.319	-0.226
	(0.144)	(0.145)	(0.055)	(0.144)	(0.105)	(0.150)	(0.128)	(0.218)	(0.232)
Median age	-0.027	0.008	0.066**	0.038	0.026	0.068	-0.038	0.016	0.142*
	(0.058)	(0.054)	(0.021)	(0.038)	(0.036)	(0.038)	(0.046)	(0.062)	(0.070)
Proportion of male	0.062	-0.118*	-0.019	0.047	0.005	-0.007	-0.026	0.083	0.109
	(0.054)	(0.052)	(0.019)	(0.046)	(0.034)	(0.043)	(0.045)	(0.067)	(0.070)
Proportion of white	-0.088	-0.258**	-0.226***	-0.456***	-0.124**	-0.262***	0.003	-0.126	-0.194
	(0.095)	(0.082)	(0.028)	(0.069)	(0.046)	(0.067)	(0.074)	(0.105)	(0.111)
Proportion of bachelor's degree or higher	0.165	-0.016	0.070*	-0.022	0.055	0.008	-0.092	-0.027	-0.063
	(0.085)	(0.080)	(0.028)	(0.063)	(0.048)	(0.060)	(0.069)	(0.099)	(0.108)
Intercept	0.177*	0.156	0.265***	-0.018	0.331***	0.075	0.157*	0.369***	0.381***
	(0.088)	(0.084)	(0.028)	(0.058)	(0.039)	(0.055)	(0.071)	(0.084)	(0.098)
Observations	150	246	1651	295	684	281	385	147	98
R^2	0.278	0.156	0.284	0.365	0.186	0.208	0.145	0.237	0.363
Adjusted R^2	0.231	0.124	0.280	0.345	0.175	0.181	0.124	0.187	0.297
Residual Std. Error	0.657	0.842	0.802	0.651	0.804	0.641	0.892	0.727	0.706
F Statistic	5.985***	4.863***	72.245***	18.226***	17.063***	7.892***	7.056***	4.727***	5.562***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Table C3. Entropy by brands' price level

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.299*	0.099	0.197***	0.215***	0.346***	0.386***	-0.054	0.106	0.035
	(0.137)	(0.059)	(0.024)	(0.034)	(0.034)	(0.043)	(0.043)	(0.062)	(0.088)
Income variability	-0.155	-0.004	0.002	-0.066**	-0.024	-0.012	-0.015	-0.101*	-0.130
	(0.121)	(0.042)	(0.018)	(0.025)	(0.027)	(0.035)	(0.031)	(0.044)	(0.070)
Mobility	-0.018	-0.003	0.018	-0.005	0.006	0.004	0.131***	-0.057	0.009
	(0.142)	(0.043)	(0.020)	(0.022)	(0.030)	(0.028)	(0.037)	(0.064)	(0.075)
Local availability	0.166	0.163***	0.129***	0.079***	0.179***	0.110***	0.168***	0.161***	0.172**
	(0.100)	(0.041)	(0.015)	(0.020)	(0.019)	(0.021)	(0.027)	(0.040)	(0.056)
In NYC	-0.114	0.511***	0.179***	-0.452***	0.297***	0.161*	0.388***	-0.296**	0.118
	(0.261)	(0.106)	(0.041)	(0.056)	(0.062)	(0.075)	(0.069)	(0.108)	(0.147)
Median age	-0.077	0.083*	0.045**	-0.038	0.011	0.078***	-0.012	0.005	0.120*
	(0.091)	(0.042)	(0.016)	(0.021)	(0.022)	(0.022)	(0.027)	(0.041)	(0.057)
Proportion of male	-0.056	0.033	0.024	-0.006	0.007	0.014	0.078**	-0.034	0.013
	(0.091)	(0.040)	(0.014)	(0.020)	(0.021)	(0.022)	(0.027)	(0.039)	(0.052)
Proportion of white	0.074	-0.025	-0.029	0.066*	-0.168***	-0.100**	0.045	-0.043	-0.256**
	(0.142)	(0.064)	(0.021)	(0.028)	(0.028)	(0.032)	(0.037)	(0.058)	(0.084)
Proportion of bachelor's degree or higher	0.050	0.197**	0.273***	-0.037	-0.210***	0.125***	0.204***	-0.078	0.220**
	(0.127)	(0.060)	(0.022)	(0.031)	(0.031)	(0.036)	(0.038)	(0.058)	(0.079)
Intercept	0.237	-0.080	-0.030	0.254***	-0.009	0.015	-0.121***	0.263***	0.043
	(0.124)	(0.063)	(0.021)	(0.027)	(0.024)	(0.030)	(0.036)	(0.051)	(0.073)
Observations	178	512	3191	1980	2098	1195	1442	539	330
R^2	0.090	0.232	0.257	0.119	0.151	0.240	0.118	0.078	0.122
Adjusted R^2	0.041	0.218	0.255	0.115	0.147	0.234	0.113	0.062	0.098
Residual Std. Error	1.113	0.913	0.833	0.875	0.872	0.727	0.975	0.912	0.981
F Statistic	1.839	16.828***	122.259***	29.532***	41.309***	41.569***	21.330***	4.982***	4.960***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Appendix D. Replicating the analyses with education.

For the CBGs in our sample, the association between median household income and proportion of people with bachelor's or higher degree is 0.710 ($p < 0.001$). Like income, using a simple LASSO regression model, we can predict proportion of people who have bachelor's or higher degree of the CBGs with the proportion of outgoing visitors of the CBGs to the brands, obtaining an out-of-sample correlation between predicted and actual proportion of people who have bachelor's degree or higher of 0.792 ($p < 0.001$, Figure D1). The distribution of the median proportion of brand visitors with bachelor's or higher degree for different Yelp price levels and for some typical brands are similar with income (Figures 1 and 2 from the main text are replicated in Figures D2 and D3).

Table D1 shows the association between CBGs' proportion of brand visitors with bachelor's or higher degree and the diversity measures by industry. The results largely replicate the findings with income: there are significant associations between education and the diversity measures in most of the industries. For three industries concerning necessity goods (*Food and Beverage Stores, General Merchandise Stores, Gasoline Stations)* the associations between education and some measures of diversity are negative but relatively weak. Table D2 shows the associations between CBGs' proportion of brand visitors with bachelor's or higher degree and diversity in consumption by CBGs in and outside New York City, showing the same trends as income.

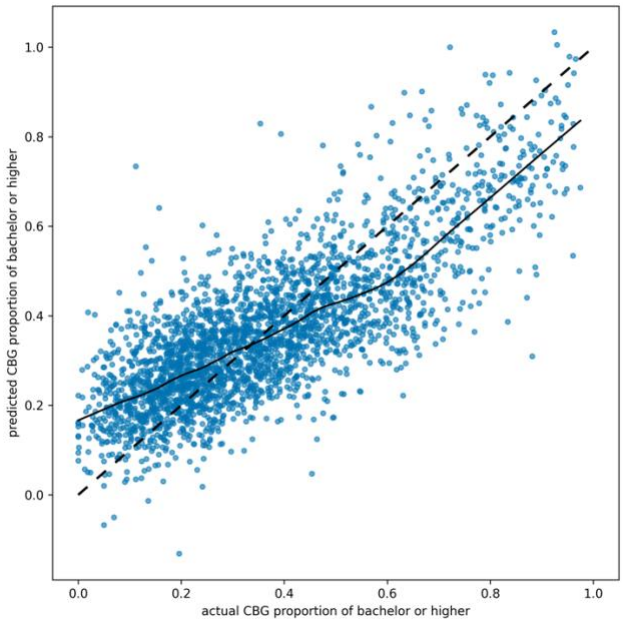

Figure D1. Correlation between predicted and actual education in the test set. The solid line shows the LOESS fit and the dashed line – the ideal 1:1 relation.
Note: Correlation 0.792 ($p < 0.001$); 0.655 variance explained.

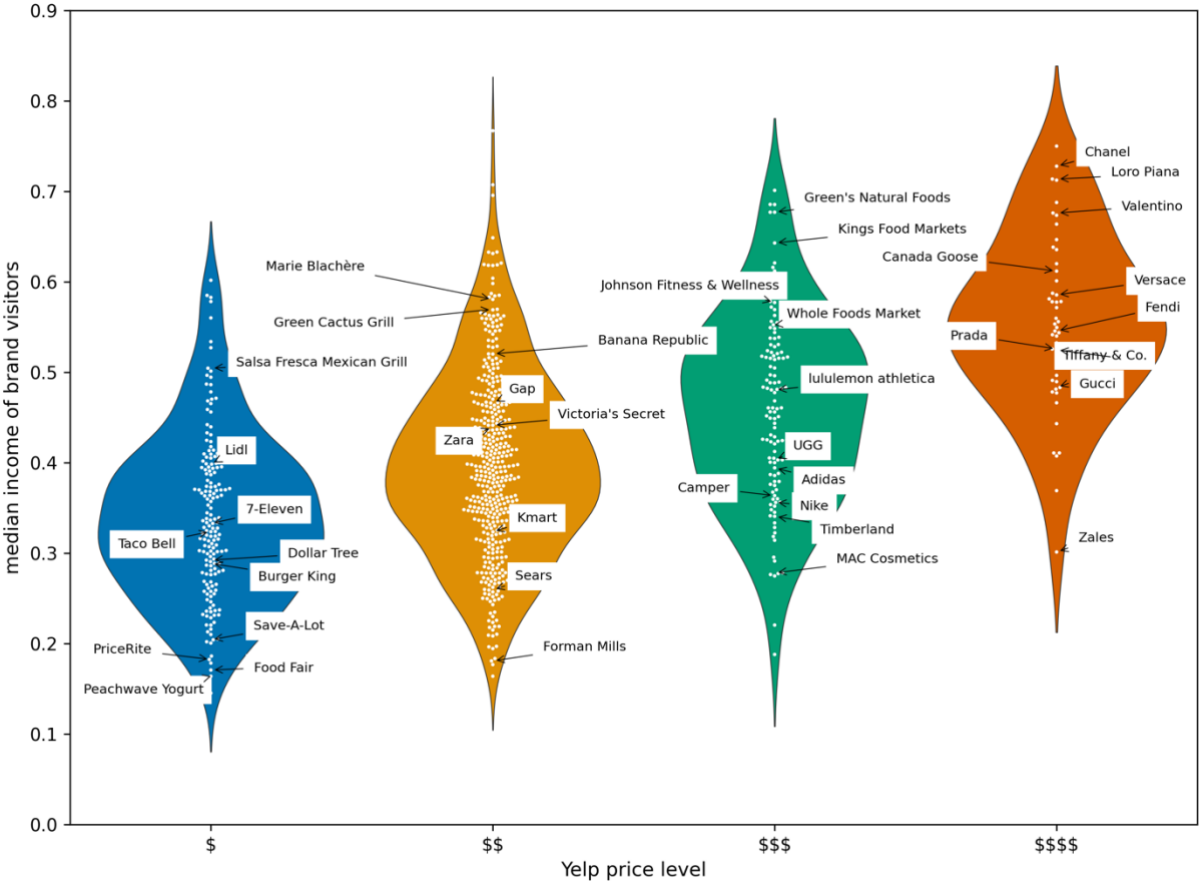

Figure D2. Relation between brands' Yelp price level and median proportion of brand visitors with bachelor's or higher degree.
Note: One extreme outlier *Balduccis* (proportion of bachelor's or higher degree 0.863; Yelp price level $$$) is excluded for better visualisation.

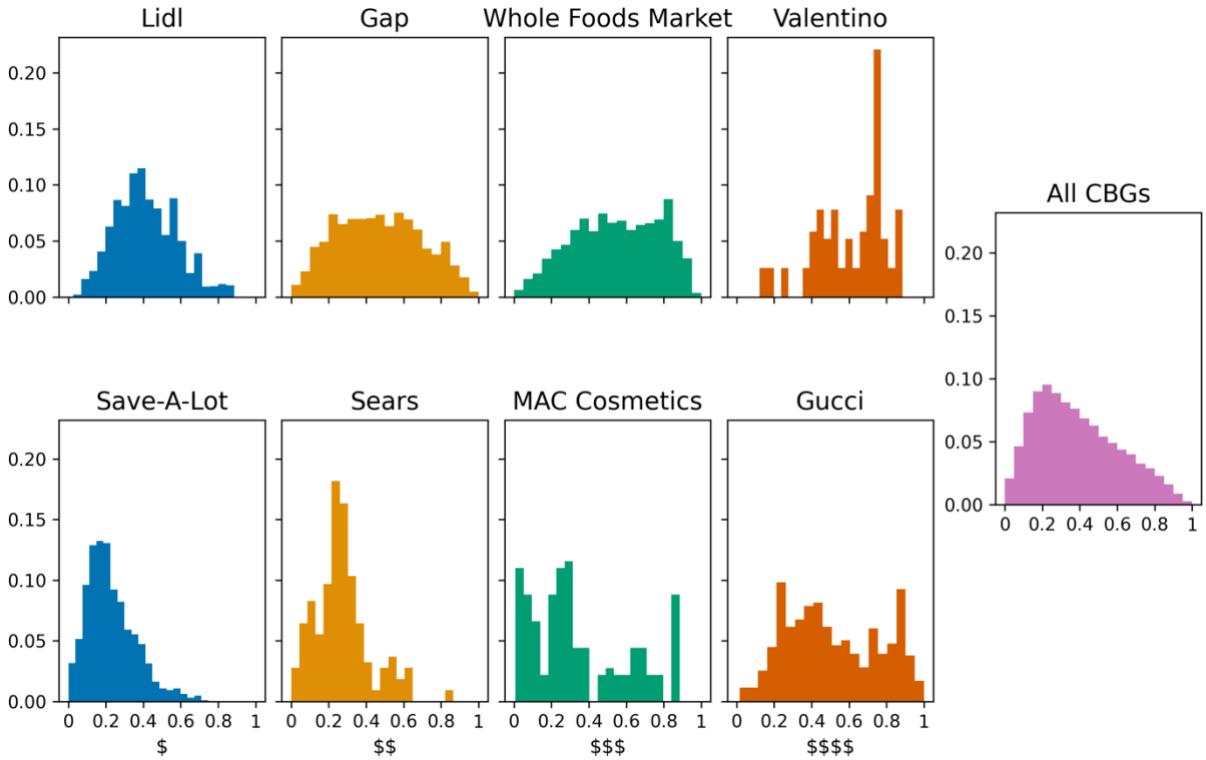

Figure D3. Distribution of proportion of brand visitors with bachelor's or higher degree for several typical brands and all CBGs in our sample.

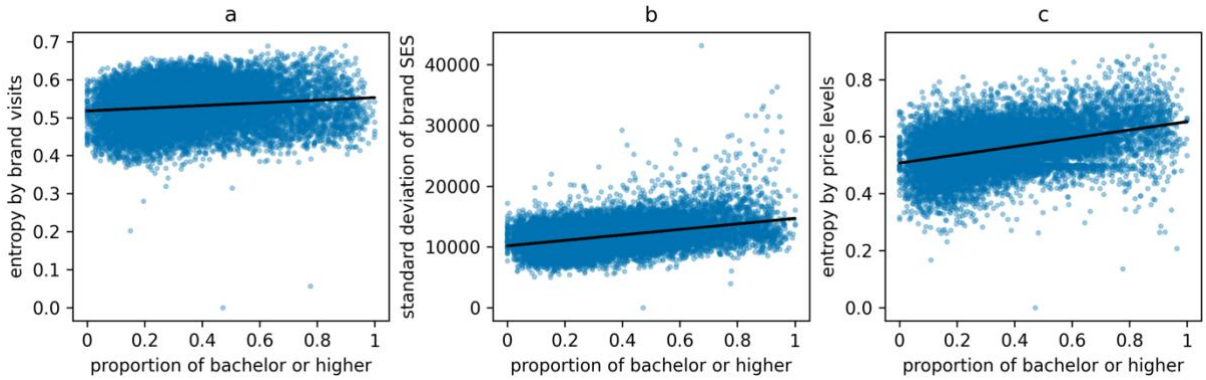

Figure D4. Correlation between CBGs' proportion of brand visitors with bachelor's or higher degree and three measures of consumption diversity.

Table D1. The associations between CBGs' proportion of residents with bachelor's or higher degree and diversity in consumption by industry.

Industry	Associations with education			Industry characteristics		
	Entropy by brand	Std of brands' SES	Entropy by brands' price level	Number of brands	Std of brands' SES	Std of brands' price level
Amusement, Gambling, and Recreation Industries	0.278***	0.218***	0.082***	89	19314	1.035
Clothing and Clothing Accessories Stores	0.142***	0.049***	0.223***	203	19654	0.768
Food Services and Drinking Places	0.141***	0.347***	0.430***	297	18584	0.663
Miscellaneous Store Retailers	0.130***	0.100***	0.013	49	23878	0.598
Sporting Goods, Hobby, Musical Instrument, and Book Stores	0.107***	0.143***	0.079***	53	25145	0.592
Motion Picture and Video Industries	0.046*	0.015	-	9	21571	-
Health and Personal Care Stores	0.034***	0.084***	0.089***	47	14221	0.494
Food and Beverage Stores	-0.064***	0.035***	-0.001	95	30776	0.640
Gasoline Stations	-0.120***	-0.090***	-0.071***	35	20154	0.512
General Merchandise Stores	-0.154***	0.055***	0.112***	42	21078	0.935

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests). Industries are ordered in descending order by entropy by brand. Yelp price levels are not available for brands in *Motion Picture and Video Industries.* Excluded *Personal and Laundry Services* and *Rental and Leasing Services* due to limited number of brands.

Table D2. The associations between CBGs' proportion of brand visitors with bachelor's or higher degree and diversity in consumption by CBGs in and outside New York City (NYC).

Region		Entropy by brand	Standard deviation of brands' SES	Entropy by brands' price level
NYC CBGs	visiting all stores	-0.155***	0.171***	0.262***
	visiting only NYC stores	0.287***	0.089***	0.103***
Non-NYC CBGs	visiting all stores	0.309***	0.494***	0.498***
	visiting only non-NYC stores	0.211***	0.490***	0.470***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).

Table D3. Results from regression analyses that predict CBGs' diversity in consumption with median household income, proportion of bachelor's degree or higher, the interaction between income and education, income variability, estimated mobility, estimated local availability, and demographic variables.

	Entropy by brand	Standard deviation of brands' SES	Entropy by brands' price level
Income	0.698***	0.403***	0.263***
	(0.019)	(0.022)	(0.019)
Proportion of bachelor's degree or higher	-0.145***	0.153***	0.218***
	(0.016)	(0.018)	(0.016)
Income-education interaction	-0.208***	-0.066***	-0.061***
	(0.010)	(0.012)	(0.010)
Income variability	-0.145***	0.039*	0.017
	(0.014)	(0.015)	(0.014)
Mobility	-0.199***	-0.164***	-0.082***
	(0.013)	(0.015)	(0.013)
Local availability	0.167***	0.228***	0.089***
	(0.011)	(0.013)	(0.011)
In NYC	-0.040	-0.466***	0.522***
	(0.030)	(0.035)	(0.030)
Median age	0.019	0.022	0.051***
	(0.012)	(0.013)	(0.012)
Proportion of male	-0.033**	-0.024*	0.007
	(0.011)	(0.012)	(0.011)
Proportion of white	-0.193***	-0.337***	-0.125***
	(0.015)	(0.018)	(0.015)
Intercept	0.165***	0.216***	-0.140***
	(0.016)	(0.018)	(0.016)
Observations	6088	3672	5739
R^2	0.295	0.406	0.319
Adjusted R^2	0.294	0.405	0.318
Residual Std. Error	0.840	0.742	0.818
F Statistic	254.038***	250.596***	268.551***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).

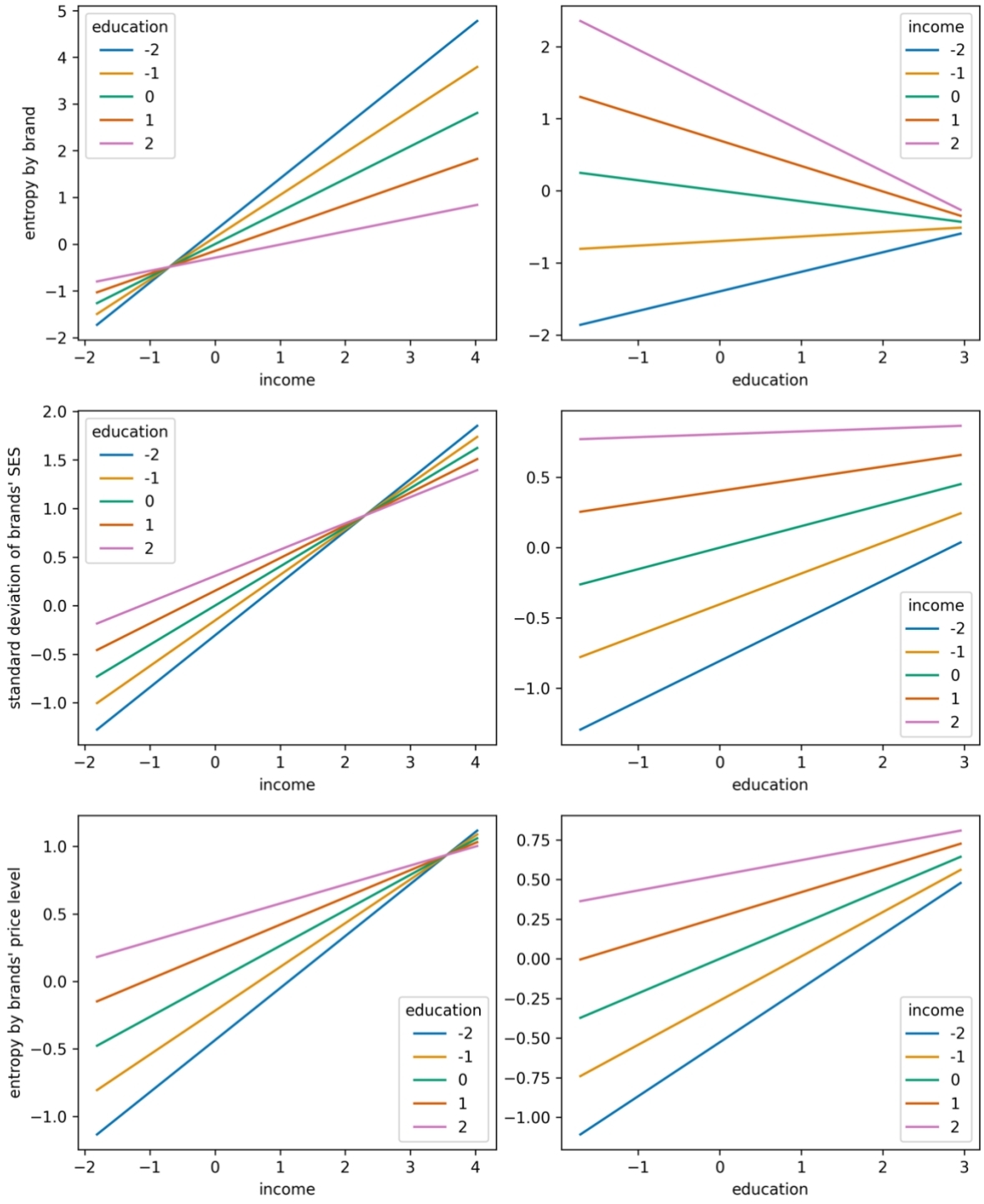

Figure D5. The interaction effects between income and education on consumption diversity.

Note: The values are standardised. For example, -2 means two standard deviations lower than the mean value. The range of income and education on the x-axis are the actual ranges from the data.

Table D4. Results from regression analyses that predict CBGs' diversity (the entropy by brand) by industry.

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.533***	0.521***	0.543***	0.374***	0.494***	0.377***	0.285***	0.519***	0.454***
	(0.051)	(0.070)	(0.026)	(0.035)	(0.026)	(0.040)	(0.043)	(0.057)	(0.085)
Proportion of bachelor's degree or higher	0.093*	-0.007	-0.079***	-0.152***	-0.235***	-0.092**	-0.075*	-0.029	0.109
	(0.042)	(0.060)	(0.022)	(0.031)	(0.024)	(0.033)	(0.035)	(0.051)	(0.067)
Income-education interaction	-0.134***	-0.134***	-0.160***	-0.067***	-0.148***	-0.203***	-0.033	-0.180***	-0.183***
	(0.027)	(0.035)	(0.014)	(0.019)	(0.017)	(0.023)	(0.022)	(0.030)	(0.042)
Income variability	-0.119***	-0.117**	-0.107***	-0.095***	-0.084***	-0.051	-0.088**	-0.137***	-0.231***
	(0.034)	(0.042)	(0.018)	(0.025)	(0.021)	(0.033)	(0.029)	(0.038)	(0.060)
Mobility	0.013	-0.074	-0.104***	0.046*	0.123***	-0.038	0.171***	0.106	0.021
	(0.051)	(0.042)	(0.020)	(0.022)	(0.023)	(0.026)	(0.034)	(0.056)	(0.067)
Local availability	0.145***	0.206***	0.192***	0.029	0.110***	0.065**	0.155***	0.131***	0.106*
	(0.030)	(0.040)	(0.015)	(0.020)	(0.015)	(0.020)	(0.024)	(0.035)	(0.048)
In NYC	0.314***	0.558***	0.190***	-0.130*	-0.417***	-0.462***	0.345***	-0.450***	-0.042
	(0.080)	(0.111)	(0.042)	(0.058)	(0.049)	(0.069)	(0.065)	(0.099)	(0.134)
Median age	0.006	0.060	-0.011	-0.004	-0.003	0.025	-0.047	0.000	0.150**
	(0.030)	(0.041)	(0.016)	(0.021)	(0.016)	(0.021)	(0.025)	(0.036)	(0.047)
Proportion of male	-0.017	0.019	-0.025	-0.029	-0.035*	-0.041*	-0.048	-0.006	0.073
	(0.028)	(0.039)	(0.015)	(0.020)	(0.016)	(0.020)	(0.025)	(0.034)	(0.044)
Proportion of white	-0.104*	-0.147*	-0.169***	-0.149***	-0.141***	-0.099***	-0.033	-0.049	-0.168*
	(0.045)	(0.063)	(0.021)	(0.028)	(0.022)	(0.030)	(0.034)	(0.050)	(0.072)
Intercept	0.215***	0.132*	0.159***	0.184***	0.427***	0.390***	0.004	0.569***	0.449***
	(0.043)	(0.064)	(0.022)	(0.029)	(0.021)	(0.031)	(0.035)	(0.045)	(0.066)
Observations	814	523	3192	2071	2559	1217	1578	608	362
R^2	0.286	0.237	0.223	0.064	0.187	0.175	0.098	0.232	0.181
Adjusted R^2	0.277	0.222	0.221	0.059	0.183	0.168	0.093	0.219	0.158
Residual Std. Error	0.816	0.914	0.840	0.891	0.752	0.677	0.942	0.834	0.894
F Statistic	32.213***	15.917***	91.408***	14.074***	58.456***	25.497***	17.111***	18.040***	7.774***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Table D5. Results from regression analyses that predict CBGs' diversity (standard deviation of brands' SES) by industry.

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.365**	0.273**	0.354***	0.276***	0.016	0.229**	0.311***	0.229*	0.451***
	(0.111)	(0.094)	(0.035)	(0.069)	(0.055)	(0.073)	(0.086)	(0.103)	(0.126)
Proportion of bachelor's degree or higher	0.214*	-0.015	0.077**	-0.022	0.060	0.011	-0.089	-0.027	0.002
	(0.083)	(0.079)	(0.028)	(0.063)	(0.049)	(0.060)	(0.070)	(0.101)	(0.114)
Income-education interaction	-0.196***	-0.082	-0.108***	0.000	-0.033	-0.041	-0.032	0.002	-0.158
	(0.057)	(0.047)	(0.018)	(0.042)	(0.037)	(0.043)	(0.037)	(0.062)	(0.098)
Income variability	0.028	0.016	0.065**	-0.032	-0.060	0.110	-0.092	0.152*	-0.101
	(0.060)	(0.057)	(0.024)	(0.048)	(0.044)	(0.061)	(0.055)	(0.068)	(0.082)
Mobility	0.252*	-0.061	-0.034	0.092	0.035	0.030	0.208**	0.053	0.405**
	(0.126)	(0.069)	(0.027)	(0.065)	(0.056)	(0.059)	(0.076)	(0.126)	(0.122)
Local availability	0.140*	0.125*	0.278***	0.264***	0.368***	0.091*	0.272***	0.176**	0.354***
	(0.056)	(0.058)	(0.020)	(0.040)	(0.031)	(0.039)	(0.049)	(0.064)	(0.075)
In NYC	0.265	0.224	-0.325***	-0.223	-0.294**	0.024	0.075	-0.324	-0.130
	(0.144)	(0.150)	(0.056)	(0.147)	(0.105)	(0.160)	(0.132)	(0.252)	(0.238)
Median age	-0.024	0.006	0.058**	0.038	0.026	0.066	-0.038	0.016	0.143*
	(0.056)	(0.054)	(0.021)	(0.038)	(0.036)	(0.039)	(0.046)	(0.062)	(0.069)
Proportion of male	0.054	-0.119*	-0.021	0.047	0.003	-0.009	-0.031	0.083	0.106
	(0.052)	(0.052)	(0.019)	(0.046)	(0.034)	(0.043)	(0.045)	(0.067)	(0.069)
Proportion of white	-0.116	-0.270**	-0.220***	-0.456***	-0.124**	-0.253***	0.007	-0.126	-0.181
	(0.092)	(0.082)	(0.028)	(0.069)	(0.046)	(0.068)	(0.074)	(0.107)	(0.111)
Intercept	0.182*	0.188*	0.311***	-0.018	0.346***	0.091	0.170*	0.369***	0.423***
	(0.085)	(0.086)	(0.028)	(0.060)	(0.042)	(0.058)	(0.073)	(0.085)	(0.101)
Observations	150	246	1651	295	684	281	385	147	98
R^2	0.334	0.167	0.299	0.365	0.187	0.210	0.147	0.237	0.381
Adjusted R^2	0.286	0.132	0.295	0.343	0.174	0.181	0.124	0.181	0.310
Residual Std. Error	0.634	0.839	0.794	0.652	0.804	0.642	0.892	0.730	0.699
F Statistic	6.970***	4.715***	69.942***	16.346***	15.434***	7.194***	6.421***	4.223***	5.356***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Table D6. Results from regression analyses that predict CBGs' diversity (standard deviation of brands' SES) by industry.

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.332*	0.179*	0.267***	0.226***	0.366***	0.417***	-0.123**	0.120	0.075
	(0.157)	(0.071)	(0.026)	(0.035)	(0.035)	(0.042)	(0.046)	(0.066)	(0.099)
Proportion of bachelor's degree or higher	0.050	0.196**	0.283***	-0.035	-0.206***	0.112**	0.202***	-0.074	0.224**
	(0.127)	(0.060)	(0.022)	(0.031)	(0.031)	(0.036)	(0.038)	(0.058)	(0.079)
Income-education interaction	-0.034	-0.073*	-0.108***	-0.027	-0.053*	-0.160***	0.088***	-0.024	-0.043
	(0.080)	(0.036)	(0.014)	(0.019)	(0.022)	(0.025)	(0.023)	(0.035)	(0.050)
Income variability	-0.164	0.006	0.007	-0.061*	-0.026	-0.013	-0.020	-0.100*	-0.127
	(0.124)	(0.042)	(0.018)	(0.025)	(0.027)	(0.035)	(0.031)	(0.044)	(0.070)
Mobility	-0.020	-0.005	0.019	-0.003	0.006	-0.005	0.130***	-0.056	0.011
	(0.142)	(0.043)	(0.019)	(0.022)	(0.030)	(0.028)	(0.037)	(0.064)	(0.075)
Local availability	0.165	0.166***	0.130***	0.079***	0.176***	0.104***	0.155***	0.162***	0.172**
	(0.100)	(0.041)	(0.015)	(0.020)	(0.019)	(0.021)	(0.027)	(0.040)	(0.056)
In NYC	-0.070	0.576***	0.227***	-0.441***	0.286***	0.195**	0.343***	-0.277*	0.158
	(0.281)	(0.110)	(0.041)	(0.057)	(0.062)	(0.074)	(0.070)	(0.112)	(0.154)
Median age	-0.077	0.080	0.034*	-0.041	0.009	0.072**	-0.005	0.002	0.119*
	(0.091)	(0.042)	(0.016)	(0.021)	(0.022)	(0.022)	(0.027)	(0.041)	(0.057)
Proportion of male	-0.057	0.031	0.021	-0.007	0.006	0.010	0.083**	-0.037	0.011
	(0.092)	(0.040)	(0.014)	(0.020)	(0.021)	(0.021)	(0.026)	(0.039)	(0.052)
Proportion of white	0.079	-0.033	-0.026	0.066*	-0.173***	-0.107***	0.042	-0.041	-0.255**
	(0.143)	(0.063)	(0.020)	(0.028)	(0.028)	(0.032)	(0.036)	(0.058)	(0.084)
Intercept	0.244	-0.052	0.021	0.267***	0.021	0.101**	-0.166***	0.272***	0.059
	(0.125)	(0.064)	(0.021)	(0.029)	(0.027)	(0.033)	(0.037)	(0.052)	(0.075)
Observations	178	512	3191	1980	2098	1195	1442	539	330
R^2	0.091	0.238	0.272	0.120	0.154	0.266	0.127	0.079	0.124
Adjusted R^2	0.036	0.223	0.269	0.115	0.150	0.260	0.121	0.062	0.097
Residual Std. Error	1.116	0.910	0.825	0.875	0.871	0.715	0.970	0.912	0.981
F Statistic	1.665	15.659***	118.570***	26.787***	37.862***	42.891***	20.850***	4.526***	4.535***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). No regression model for the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data.

Appendix E. Replicating the analyses with Texas data.

Data

After filtering the 28 four-digit NAICS codes, there are 323,098 observations in total for the Texas data in October 2019. Table E1 shows the four-digit category names, the NAICS codes, and the number of POIs for each category. Among the 323,098 observations, 83,299 (about 26 percent) have brands associated with them. After dropping observations with limited information, we have 56,266 POIs that belong to 1,484 brands. Among the 28 categories, two of them (*Gambling Industries* and *Museums, Historical Sites, and Similar Institutions*), do not have any brand, so we end up with 26 categories. For 55,268 of the 56,266 POIs, each observation has a table of the visitors' home CBGs and the number of visitors from each CBG. The visitors to the 55,268 POIs are from 66,523 CBGs, where 64,753 median household income estimates are available. Aggregating the data by brands, we get a bipartite network of 64,753 CBGs and 1,459 brands, where the weights are the number of visits from each CBG to each brand. We drop the brands that have fewer than 100 incoming visitors and CBGs that have fewer than 100 outgoing visitors, resulting in a network of 15,729 CBGs and 1,273 brands. As shown in Figure E1, compared with all 64,753 CBGs, the selected 15,729 CBGs slightly oversample lower income CBGs. For the 1,273 brands, we were able to match 511 brands with the Yelp Open Dataset and manually found the Yelp price level for 762 brands, leaving 265 brands missing.

Table E1. Four-digit NAICS industries and number of POIs included in the Texas data.

Four-digit industry name	NAIC codes	Number of POIs
Restaurants and Other Eating Places	7225	97658
Personal Care Services	8121	55192
Gasoline Stations	4471	19588
Other Amusement and Recreation Industries	7139	17889
Grocery Stores	4451	17377
Health and Personal Care Stores	4461	15364
Clothing Stores	4481	12814
Museums, Historical Sites, and Similar Institutions	7121	11073
Other Miscellaneous Store Retailers	4539	9237
Drinking Places (Alcoholic Beverages)	7224	8215
Sporting Goods, Hobby, and Musical Instrument Stores	4511	7884
Consumer Goods Rental	5322	7586
Specialty Food Stores	4452	5900
Jewelry, Luggage, and Leather Goods Stores	4483	5872
General Merchandise Stores, including Warehouse Clubs and Supercenters	4523	5477
Florists	4531	4098
Used Merchandise Stores	4533	4075
Beer, Wine, and Liquor Stores	4453	4036
Office Supplies, Stationery, and Gift Stores	4532	3362
Shoe Stores	4482	2587
Amusement Parks and Arcades	7131	1614
Special Food Services	7223	1508
Department Stores	4522	1495
Book Stores and News Dealers	4512	1387
Gambling Industries	7132	693
Motion Picture and Video Industries	5121	692
Spectator Sports	7112	232
Performing Arts Companies	7111	193

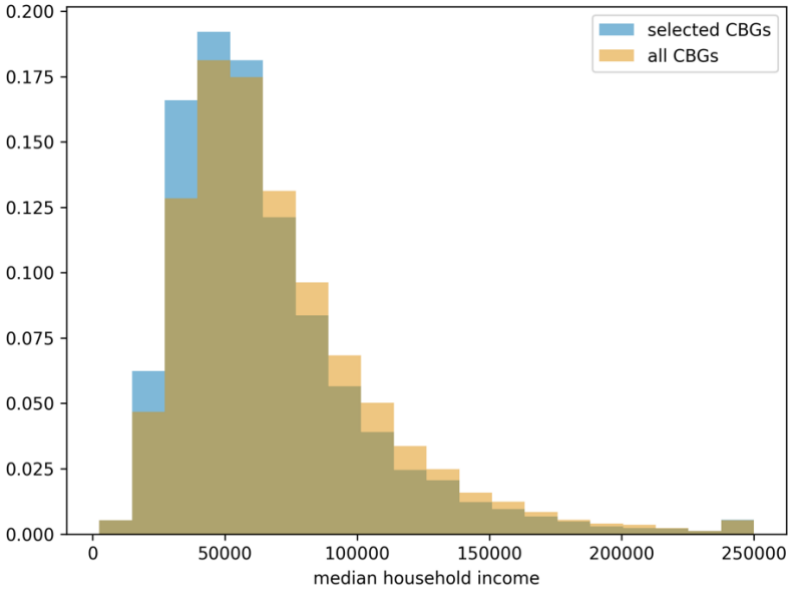

Figure E1. The income distribution for selected and all CBGs in the Texas data.

Descriptive results

Using simple LASSO regression models, we can predict the median household income or the proportion of people who have bachelor's or higher degree of the CBGs with the proportion of outgoing visitors of the CBGs to the brands. The out-of-sample correlations are 0.753 for income and 0.859 for education, both statistically significant at the 0.001 level. The models explain 0.640 variance for income and 0.767 variance for education. Figure E2 shows the correlation between predicted and actual income (panel a) or education (panel b) in the test set.

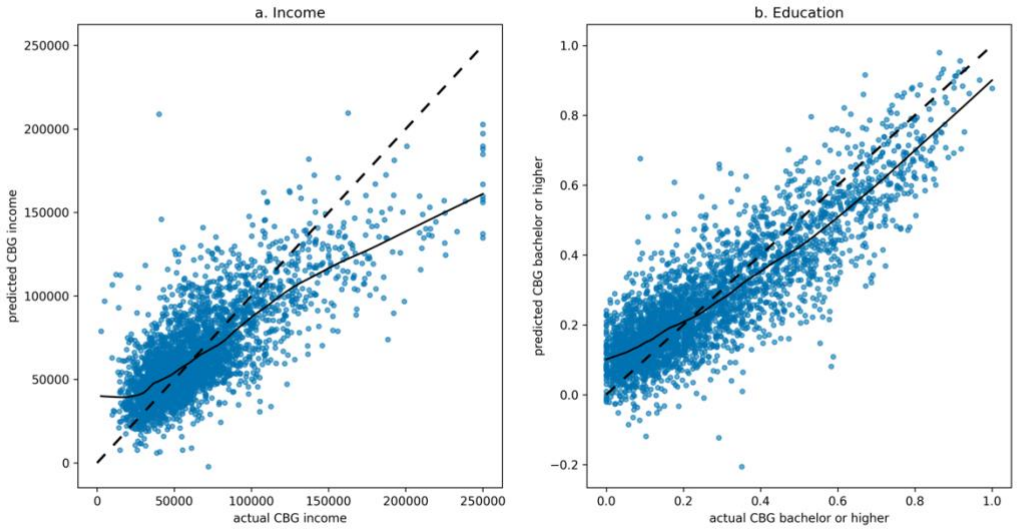

Figure E2. Correlation between predicted and actual income or education in the test set (Texas). The solid lines show the LOESS fit and the dashed lines – the ideal 1:1 relation.

Figure E3 shows the distribution of brands' SES (measured by the median income of brand visitors) for different Yelp price levels, with some typical brands labelled. Figure E4 shows the distribution of brand visitors' income for some typical brands identified from Figure E3, and for reference, the distribution of median household income for the CBGs in our sample. We try to use the same brands as in New York State, but Lidl and Valentino do not have data available in Texas, so they are replaced with similar brands 7-Eleven and Jimmy Choo. The patterns are similar with what we find in New York State. Similar patterns exist for education, as shown in Figure E5 and E6.

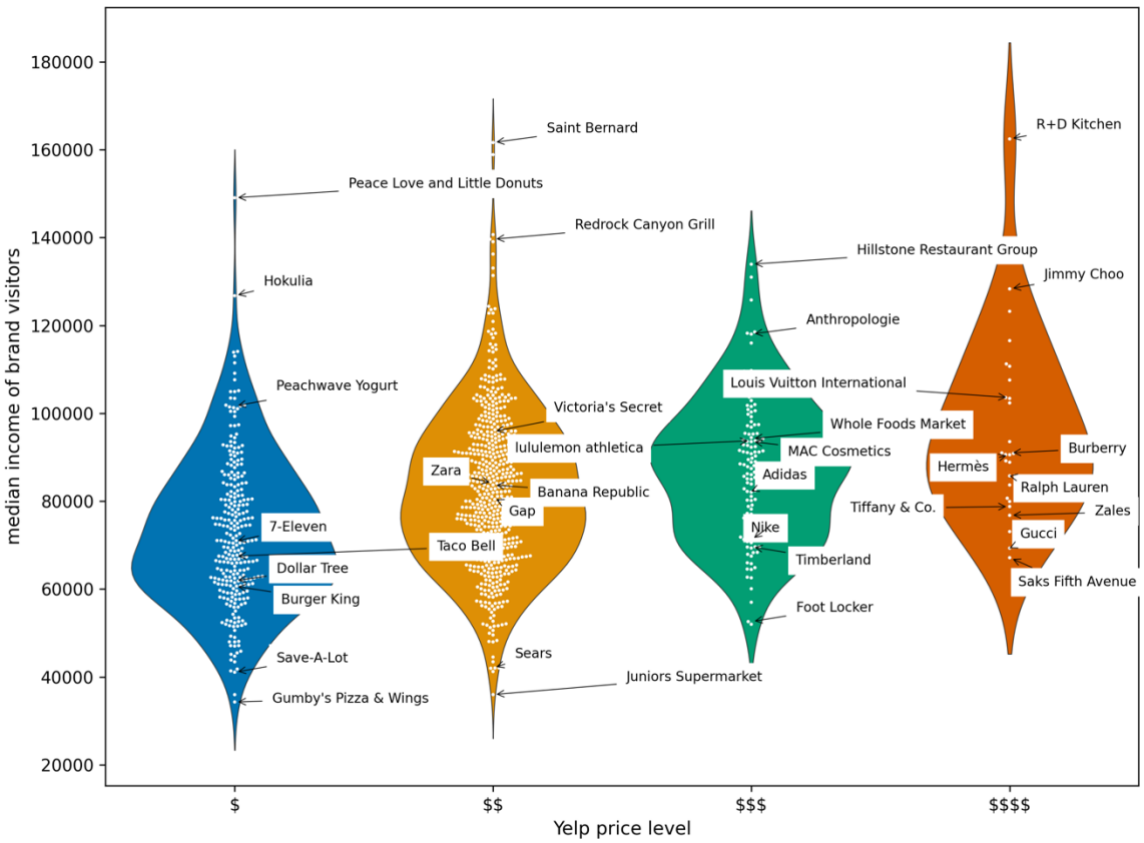

Figure E3. Relation between brands' Yelp price level and median income of brand visitors (Texas). Note: One extreme outlier *The Pizza Press* (median income 9,222; yelp price level $) is excluded for better visualisation.

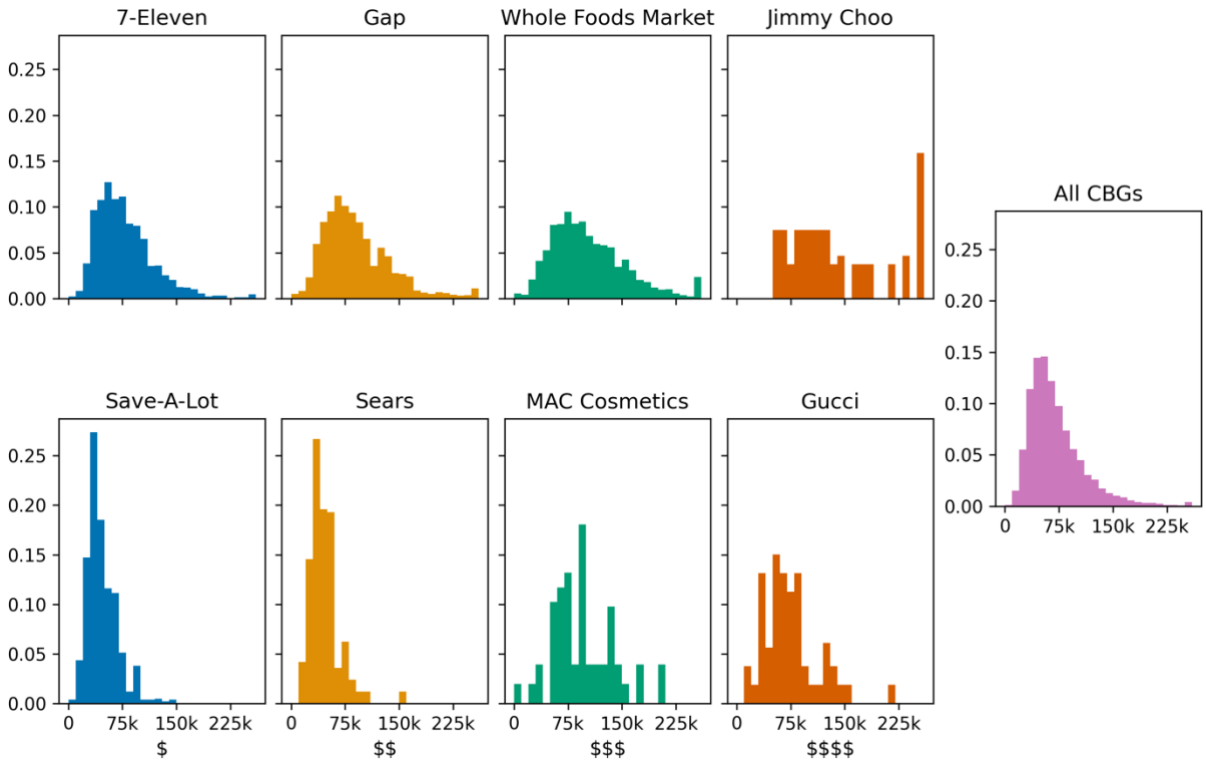

Figure E4. Distribution of brand visitors' income for some typical brands and all CBGs in our sample (Texas).

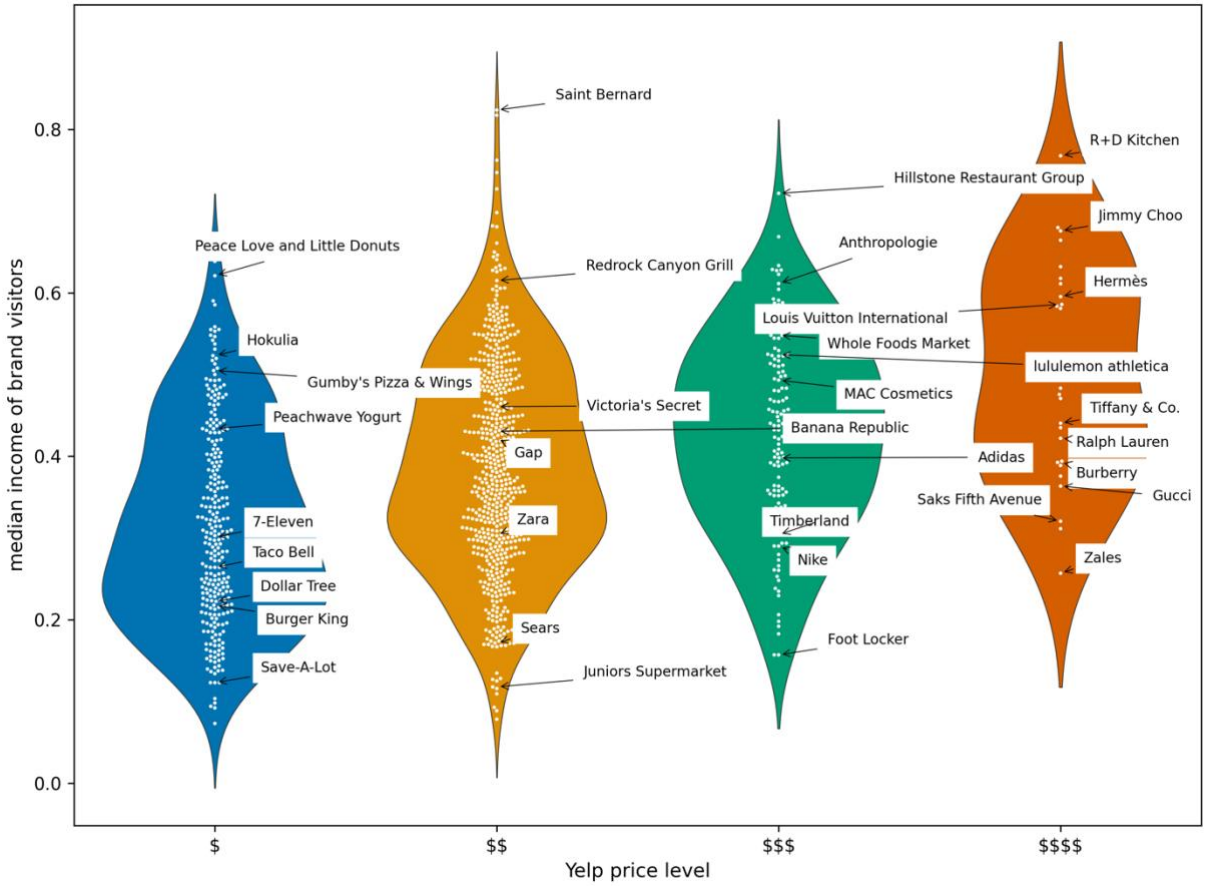

Figure E5. Relation between brands' Yelp price level and median proportion of brand visitors with bachelor's or higher degree (Texas).

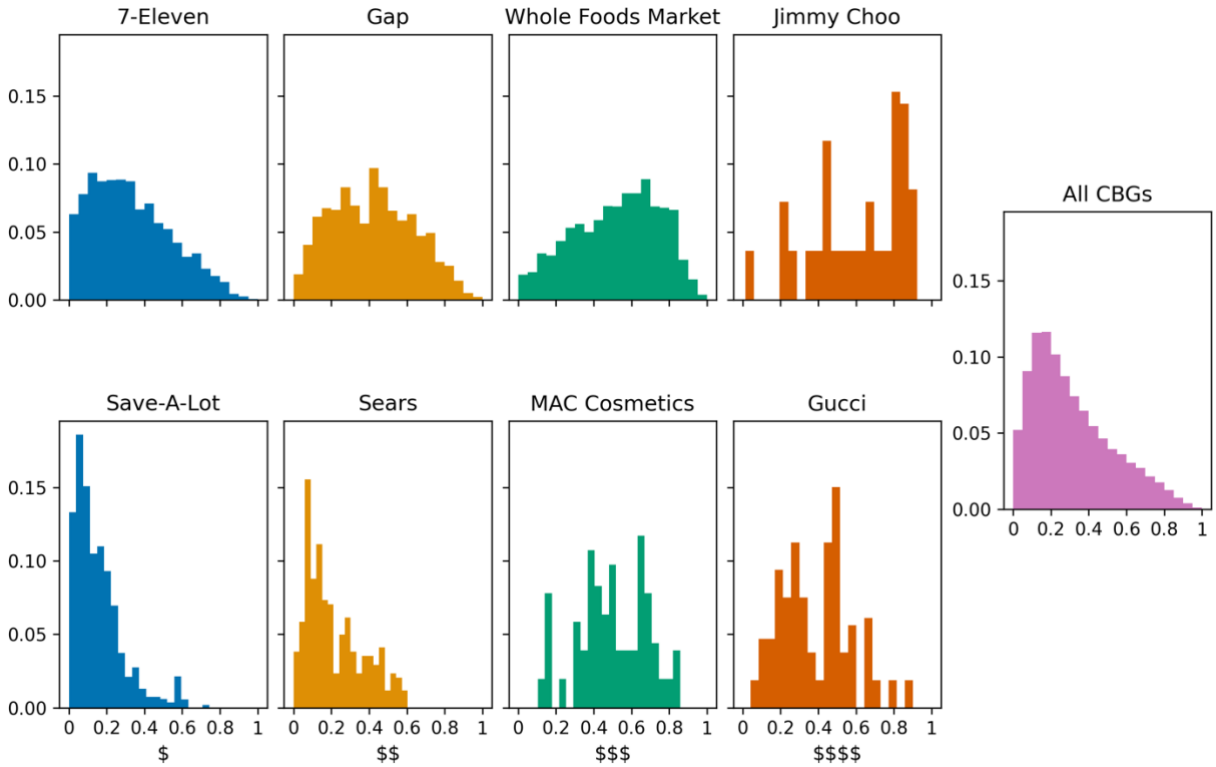

Figure E6. Distribution of proportion of brand visitors with bachelor's or higher degree for typical brands and all CBGs in our sample (Texas).

Diversity in consumption

The consumption diversity hypothesis is confirmed again with Texas data. As shown in Table E2, all measures of diversity have significant and positive correlation with CBGs' median household income. Figure E7 shows the correlations between CBGs' median household income and the three measures of diversity that we focus on reporting. The correlation coefficients are 0.433, 0.560, and 0.473 respectively for the entropy by brand, the standard deviation of visited brands' SES, and the entropy by brands' price level. Figure E8 shows correlation between CBGs' proportion of residents with bachelor's or higher degree and the three measures of diversity. The correlation coefficients are 0.395, 0.663, and 0.557 respectively. Notably, the correlations in Texas are stronger than in New York State, especially for education, which we interpret to indicate higher consumption inequality in Texas.

Table E2. Correlations between median household income, proportion of residents who have a bachelor's or higher degree, and six different measures of diversity in consumption (Texas).

	Income	**Proportion of bachelor or higher**	**Number of brands**	**Brand entropy**	**SES range**	**SES std**	**Number of price level**	**Price entropy**
Income	1	0.732***	0.409***	0.433***	0.432***	0.560***	0.322***	0.473***
Proportion of bachelor or higher	0.732***	1	0.331***	0.395***	0.443***	0.663***	0.318***	0.557***
Number of brands	0.409***	0.331***	1	0.850***	0.603***	0.344***	0.530***	0.298***
Brand entropy	0.433***	0.395***	0.85***	1	0.649***	0.479***	0.547***	0.395***
SES range	0.432***	0.443***	0.603***	0.649***	1	0.727***	0.432***	0.372***
SES std	0.560***	0.663***	0.344***	0.479***	0.727***	1	0.349***	0.530***
Number of price level	0.322***	0.318***	0.530***	0.547***	0.432***	0.349***	1	0.594***
Price entropy	0.473***	0.557***	0.298***	0.395***	0.372***	0.530***	0.594***	1

Note: *** $p < 0.001$ (two-tailed tests).

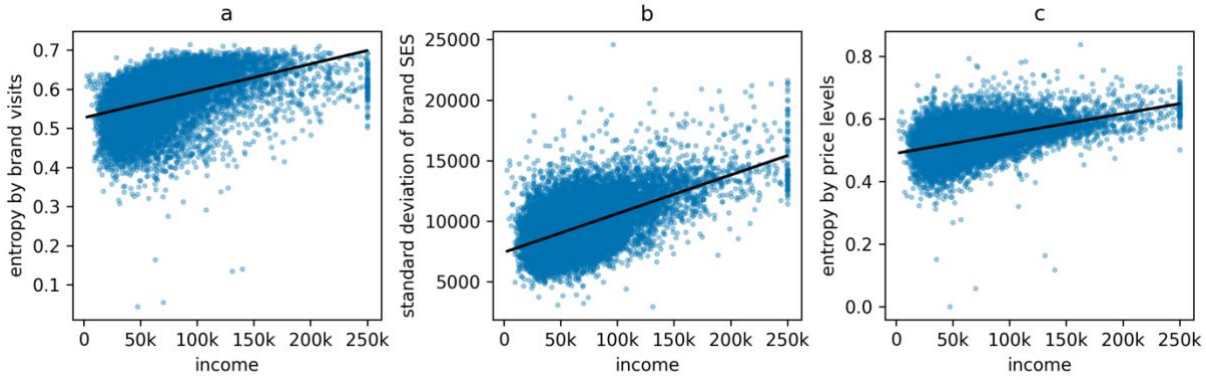

Figure E7. Correlation between CBGs' median household income and three measures of diversity (Texas).

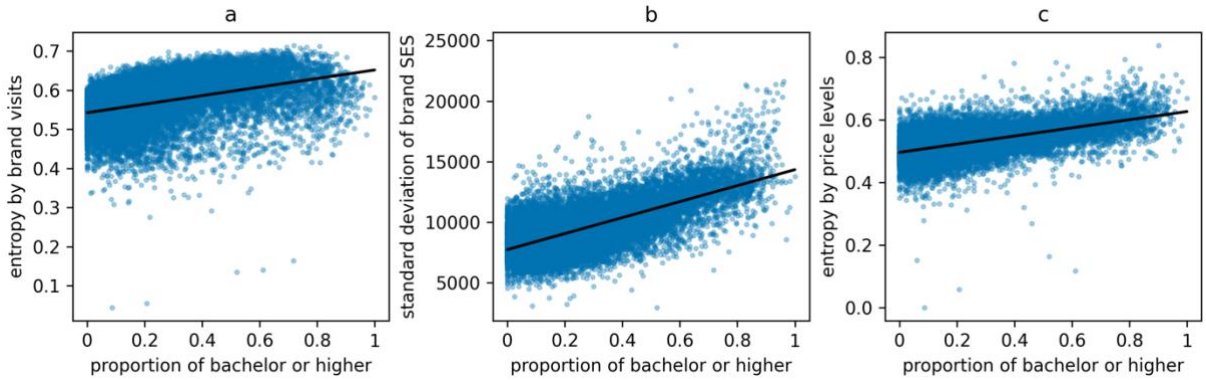

Figure E8. Correlation between CBGs' proportion of residents with bachelor's or higher degree and three measures of diversity (Texas).

Table E3 shows the association between CBGs' median household income or CBGs' proportion of residents with bachelor's or higher degree and the diversity measures by industry. With only one exception of no correlation, there are significant associations between income or education and the diversity measures in all the industries, confirming the robustness of the consumption diversity hypothesis. The association is again stronger in industries that involve more cultural aspects than those that concern necessity goods. Compared with the results for New York State, where the associations between education and some measures of diversity are negative for a few industries involving necessity goods, the results in Texas provide an even stronger validation of our hypothesis.

To explore urban-rural lifestyle differences, we disaggregate the analysis by Greater Houston and the rest of the state. We select Greater Houston because Houston is the largest and most

populous city in Texas, but the distinction between Greater Houston and the rest of Texas is not as pronounced as New York City and the rest of New York State. We study the following cases: visits from CBGs in Greater Houston to stores in Texas, visits from CBGs outside of Greater Houston to stores in Texas, visits from CBGs in Greater Houston only to stores in Greater Houston, and visits from CBGs outside of Greater Houston to stores outside of Greater Houston. Table E4 shows the associations between CBGs' income or education and the diversity measures for those cases. There are some differences in the associations between income or education and diversity between CBGs inside and outside of Greater Houston, but the differences are not as substantial as in New York City. It seems that New York City is a special case of weak association between socioeconomic status and consumption diversity.

Table E3. The associations between CBGs' median household income or CBGs' proportion residents with bachelor's or higher degree and diversity in consumption by industry.

Industry	Associations with income			Association with education			Industry characteristics		
	Entropy by brand	Std of brands' SES	Entropy by brands' price level	Entropy by brand	Std of brands' SES	Entropy by brands' price level	Number of brands	Std of brands' SES	Std of brands' price level
Amusement, Gambling, and Recreation Industries	0.445***	0.280***	0.203***	0.432***	0.301***	0.179***	116	21331	0.667
Food Services and Drinking Places	0.396***	0.514***	0.526***	0.362***	0.616***	0.624***	554	19026	0.619
Miscellaneous Store Retailers	0.344***	0.268***	0.156***	0.308***	0.249***	0.126***	61	18130	0.581
Sporting Goods, Hobby, Musical Instrument, and Book Stores	0.338***	0.237***	0.202***	0.292***	0.218***	0.260***	60	22591	0.572
Clothing and Clothing Accessories Stores	0.319***	0.177***	0.269***	0.268***	0.182***	0.259***	210	17306	0.686
Health and Personal Care Stores	0.262***	0.249***	0.023**	0.212***	0.222***	0.024**	56	14842	0.436
General Merchandise Stores	0.203***	0.453***	0.397***	0.158***	0.524***	0.472***	41	16618	0.897
Motion Picture and Video Industries	0.189***	0.241***	-	0.165***	0.21***	-	14	24893	-
Food and Beverage Stores	0.177***	0.165***	0.115***	0.136***	0.214***	0.142***	98	18558	0.634
Personal and Laundry Services	0.139***	0.109**	0.096*	0.121***	0.081*	0.088*	16	19242	0.641
Gasoline Stations	0.136***	0.400***	0.031***	0.029***	0.400***	-0.005	40	9205	0.512

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests). Industries are ordered in descending order by the association between income and entropy by brand. Yelp price levels are not available for brands in *Motion Picture and Video Industries*. Excluded *Performing Arts, Spectator Sports, and Related Industries* and *Rental and Leasing Services* due to limited number of brands.

Table E4. The associations between CBGs' median household income and diversity in consumption by CBGs in and outside Greater Houston.

Region		Associations with income			Associations with education		
		Entropy by brand	**Std of brands' SES**	**Entropy by brands' price level**	**Entropy by brand**	**Std of brands' SES**	**Entropy by brands' price level**
Greater Houston CBGs	visiting all stores	0.392***	0.539***	0.539***	0.356***	0.707***	0.634***
	visiting only Greater Houston stores	0.174***	0.421***	0.376***	0.272***	0.517***	0.455***
Non-Greater Houston CBGs	visiting all stores	0.441***	0.577***	0.451***	0.406***	0.681***	0.539***
	visiting only Non-Greater Houston stores	0.437***	0.575***	0.452***	0.401***	0.679***	0.539***

Note: *** $p < 0.001$ (two-tailed tests).

Robustness to alternative explanations

As in New York State, we use a standardised regression model to test the extent to which the associations between high SES and consumption diversity can be explained by simple geographic constraints. Table E5 shows the results from the regression models using different diversity measures. For all three measures of diversity, income and education are significantly associated with diversity, even controlling for income variability, mobility, local availability, age, gender, and race. Income is the predominant predictor for the entropy by brand, while education is the primary predictor for the standard deviation of brand's SES and the entropy by brands' price level. There is a significantly negative effect of the interaction between income and education for the entropy by brand. As shown in Figure E9, the negative effect means that the association between income and entropy by brand visits is weaker for CBGs with higher education and the association between education and entropy by brand visits is negative for CBGs with higher income. There is a significantly positive but relatively weak effect of the interaction between income and education for the standard deviation of brands' SES, meaning the association is stronger for the high income and high education group. The interaction effect is not significant for the entropy by brands' price level.

The findings are similar if we repeat the regression analyses separately by industry (see Tables E6, E7, E8). These results affirm the central hypothesis of this paper that high SES is associated with diverse consumption, but they also indicate nuances in the association that should be explored by future research with more refined data.

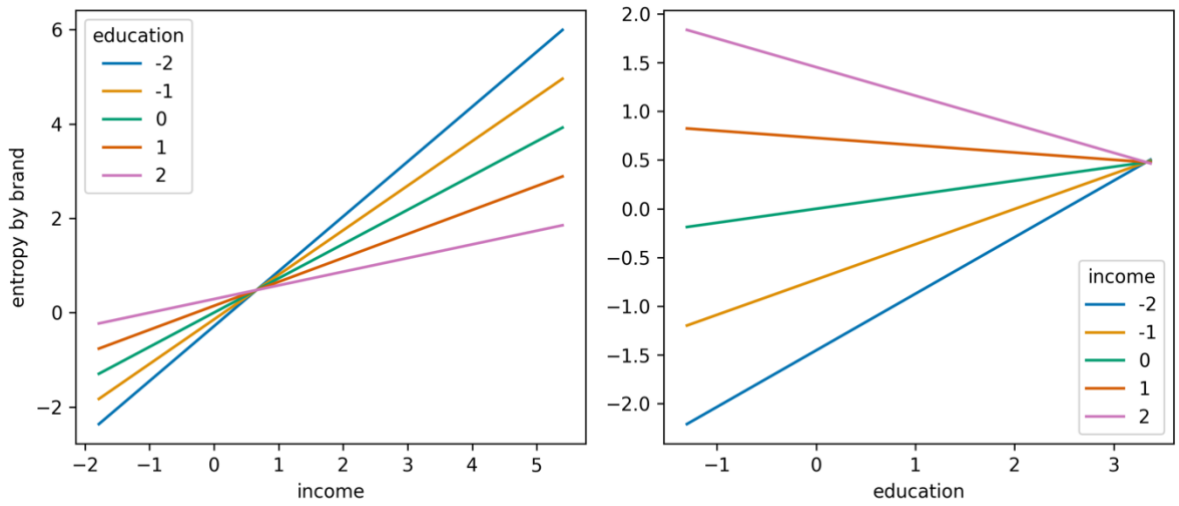

Figure E9. The interaction effects between income and education on the entropy by brand (Texas).

Note: The values are standardised. For example, -2 means two standard deviations lower than the mean value. The ranges of income and education on the x-axes are the actual ranges from the data.

Table E5. Results from regression analyses that predict CBGs' diversity in consumption with median household income, proportion of bachelor's degree or higher, the interaction between income and education, income variability, estimated mobility, estimated local availability, and demographic variables (Texas).

	Entropy by brand	Standard deviation of brands' SES	Entropy by brands' price level
Income	0.726***	0.085***	0.127***
	(0.014)	(0.014)	(0.015)
Proportion of bachelor's degree or higher	0.144***	0.498***	0.461***
	(0.011)	(0.011)	(0.012)
Income-education interaction	-0.218***	0.064***	-0.014
	(0.007)	(0.007)	(0.007)
Income variability	-0.173***	0.005	-0.003
	(0.009)	(0.009)	(0.010)
Mobility	-0.186***	-0.221***	-0.063***
	(0.008)	(0.009)	(0.009)
Local availability	0.140***	0.110***	0.110***
	(0.008)	(0.008)	(0.008)
In Greater Houston	0.274***	0.436***	0.307***
	(0.019)	(0.019)	(0.021)
Median age	-0.103***	-0.012	0.024**
	(0.008)	(0.008)	(0.009)
Proportion of male	-0.029***	0.017*	0.015
	(0.007)	(0.007)	(0.008)
Proportion of white	-0.050***	-0.093***	0.014
	(0.008)	(0.008)	(0.009)
Intercept	0.099***	-0.131***	-0.051***
	(0.010)	(0.010)	(0.011)
Observations	9571	7202	9007
R^2	0.479	0.611	0.375
Adjusted R^2	0.479	0.610	0.374
Residual Std. Error	0.722	0.628	0.784
F Statistic	879.770***	1128.589***	540.136***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).

Table E6. Results from regression analyses that predict CBGs' diversity (entropy by brand) by industry (Texas).

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.669***	0.524***	0.693***	0.333***	0.569***	0.354***	0.434***	0.547***	0.496***
	(0.032)	(0.054)	(0.017)	(0.025)	(0.020)	(0.028)	(0.033)	(0.040)	(0.051)
Proportion of bachelor's degree or higher	0.156***	0.052	0.139***	0.001	-0.142***	0.152***	0.023	0.049	0.114**
	(0.023)	(0.039)	(0.013)	(0.019)	(0.017)	(0.024)	(0.025)	(0.029)	(0.038)
Income-education interaction	-0.192***	-0.126***	-0.215***	-0.084***	-0.209***	-0.107***	-0.132***	-0.144***	-0.138***
	(0.015)	(0.027)	(0.008)	(0.012)	(0.011)	(0.016)	(0.016)	(0.018)	(0.025)
Income variability	-0.175***	-0.173***	-0.177***	-0.056***	-0.115***	-0.115***	-0.132***	-0.162***	-0.169***
	(0.018)	(0.033)	(0.011)	(0.016)	(0.013)	(0.018)	(0.021)	(0.022)	(0.031)
Mobility	0.020	-0.027	-0.176***	0.022	0.103***	-0.190***	0.038	0.236***	-0.098**
	(0.033)	(0.032)	(0.011)	(0.016)	(0.012)	(0.016)	(0.031)	(0.054)	(0.034)
Local availability	0.143***	0.280***	0.152***	0.048***	0.148***	0.110***	0.147***	0.131***	0.135***
	(0.018)	(0.030)	(0.010)	(0.013)	(0.011)	(0.015)	(0.019)	(0.022)	(0.028)
In NYC	0.135**	0.410***	0.056*	0.684***	0.400***	0.265***	0.279***	0.184***	0.094
	(0.042)	(0.072)	(0.024)	(0.035)	(0.028)	(0.039)	(0.045)	(0.050)	(0.068)
Median age	-0.135***	-0.195***	-0.101***	-0.097***	0.002	-0.114***	-0.091***	-0.133***	-0.116***
	(0.021)	(0.033)	(0.011)	(0.015)	(0.012)	(0.015)	(0.022)	(0.026)	(0.032)
Proportion of male	-0.039*	-0.064*	-0.031***	-0.007	-0.042***	-0.007	-0.084***	-0.070**	-0.067*
	(0.019)	(0.029)	(0.009)	(0.014)	(0.011)	(0.015)	(0.020)	(0.023)	(0.030)
Proportion of white	-0.073***	-0.020	-0.052***	-0.220***	-0.145***	0.008	-0.029	-0.061*	-0.045
	(0.022)	(0.033)	(0.010)	(0.015)	(0.012)	(0.015)	(0.021)	(0.025)	(0.032)
Intercept	0.369***	0.196***	0.196***	-0.042*	0.127***	0.062**	0.171***	0.406***	0.286***
	(0.024)	(0.039)	(0.012)	(0.018)	(0.014)	(0.019)	(0.024)	(0.033)	(0.037)
Observations	1595	818	6010	3671	6214	3509	2053	1160	750
R^2	0.434	0.322	0.429	0.256	0.210	0.212	0.177	0.286	0.269
Adjusted R^2	0.430	0.314	0.428	0.254	0.209	0.210	0.173	0.280	0.259
Residual Std. Error	0.691	0.830	0.738	0.809	0.864	0.844	0.838	0.725	0.752
F Statistic	121.389***	38.366***	451.107***	126.212***	164.680***	94.320***	43.836***	46.000***	27.230***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). Excluded the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data. Excluded *Personal and Laundry Services* due to limited number of observations.

Table E7. Results from regression analyses that predict CBGs' diversity (standard deviation of brands' SES) by industry (Texas).

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.129*	0.118	0.078***	0.019	0.231***	0.233***	0.224***	0.174*	-0.194*
	(0.064)	(0.078)	(0.020)	(0.045)	(0.027)	(0.042)	(0.067)	(0.080)	(0.094)
Proportion of bachelor's degree or higher	0.118**	0.106	0.484***	0.172***	0.224***	0.351***	0.056	0.177**	0.128*
	(0.042)	(0.056)	(0.015)	(0.034)	(0.023)	(0.036)	(0.053)	(0.062)	(0.064)
Income-education interaction	-0.032	-0.015	0.057***	0.018	-0.038*	0.011	-0.060	-0.021	0.220***
	(0.031)	(0.035)	(0.010)	(0.023)	(0.015)	(0.023)	(0.034)	(0.039)	(0.042)
Income variability	-0.051	-0.014	0.024	0.036	-0.032	-0.049	-0.096*	0.023	-0.011
	(0.035)	(0.040)	(0.013)	(0.029)	(0.017)	(0.027)	(0.041)	(0.045)	(0.047)
Mobility	0.101	-0.125*	-0.229***	-0.067	-0.075***	-0.177***	0.009	0.074	-0.060
	(0.116)	(0.054)	(0.015)	(0.038)	(0.017)	(0.027)	(0.071)	(0.127)	(0.068)
Local availability	0.141***	0.012	0.134***	0.215***	0.231***	0.117***	0.244***	0.203***	0.357***
	(0.032)	(0.043)	(0.012)	(0.025)	(0.014)	(0.022)	(0.038)	(0.043)	(0.046)
In NYC	-0.107	0.080	0.290***	0.386***	0.759***	0.172**	0.711***	-0.105	-0.030
	(0.073)	(0.097)	(0.027)	(0.060)	(0.034)	(0.055)	(0.087)	(0.093)	(0.109)
Median age	0.004	-0.071	0.009	-0.072**	0.030	0.027	0.042	0.029	0.021
	(0.045)	(0.047)	(0.013)	(0.027)	(0.016)	(0.024)	(0.049)	(0.055)	(0.060)
Proportion of male	-0.029	-0.075	0.014	-0.009	-0.039**	-0.033	-0.068	-0.136**	-0.066
	(0.038)	(0.043)	(0.011)	(0.025)	(0.014)	(0.023)	(0.042)	(0.048)	(0.056)
Proportion of white	0.059	-0.058	-0.094***	-0.173***	-0.053**	-0.103***	-0.012	-0.133**	-0.102
	(0.041)	(0.046)	(0.012)	(0.028)	(0.016)	(0.023)	(0.045)	(0.047)	(0.053)
Intercept	0.407***	0.119*	-0.072***	-0.109***	-0.142***	-0.100***	0.178***	0.254***	0.096
	(0.061)	(0.056)	(0.015)	(0.032)	(0.019)	(0.028)	(0.051)	(0.071)	(0.064)
Observations	373	364	3822	1007	2510	957	540	356	241
R^2	0.155	0.097	0.559	0.272	0.427	0.434	0.258	0.258	0.400
Adjusted R^2	0.132	0.071	0.557	0.264	0.425	0.428	0.244	0.237	0.374
Residual Std. Error	0.611	0.775	0.686	0.753	0.713	0.653	0.860	0.766	0.688
F Statistic	6.664***	3.771***	482.304***	37.156***	186.485***	72.492***	18.400***	12.015***	15.354***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). Excluded the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data. Excluded *Personal and Laundry Services* due to limited number of observations.

Table E8. Results from regression analyses that predict CBGs' diversity (entropy by brands' price level) by industry (Texas).

	Amusement, Gambling, and Recreation Industries	Clothing and Clothing Accessories Stores	Food Services and Drinking Places	Food and Beverage Stores	Gasoline Stations	General Merchandise Stores	Health and Personal Care Stores	Miscellaneous Store Retailers	Sporting Goods, Hobby, Musical Instrument, and Book Stores
Income	0.566***	0.237***	0.198***	0.119***	0.182***	0.296***	0.043	0.268***	-0.003
	(0.072)	(0.065)	(0.016)	(0.029)	(0.024)	(0.026)	(0.041)	(0.050)	(0.069)
Proportion of bachelor's degree or higher	0.076	0.127**	0.515***	0.144***	-0.051*	0.448***	0.041	-0.061	0.209***
	(0.058)	(0.044)	(0.013)	(0.022)	(0.020)	(0.023)	(0.031)	(0.036)	(0.052)
Income-education interaction	-0.221***	-0.035	-0.051***	-0.076***	-0.082***	-0.122***	-0.025	-0.072**	-0.008
	(0.039)	(0.031)	(0.008)	(0.014)	(0.013)	(0.015)	(0.019)	(0.023)	(0.035)
Income variability	-0.207***	-0.048	-0.027**	0.003	-0.007	-0.067***	-0.050*	-0.107***	-0.036
	(0.047)	(0.037)	(0.010)	(0.018)	(0.016)	(0.017)	(0.025)	(0.028)	(0.043)
Mobility	-0.048	0.027	-0.131***	-0.058**	0.057***	-0.164***	0.019	0.243***	-0.028
	(0.070)	(0.038)	(0.011)	(0.018)	(0.014)	(0.015)	(0.042)	(0.066)	(0.045)
Local availability	0.086*	0.201***	0.126***	0.085***	0.118***	0.107***	0.163***	0.092***	0.181***
	(0.038)	(0.035)	(0.009)	(0.016)	(0.013)	(0.014)	(0.023)	(0.026)	(0.038)
In NYC	-0.076	0.399***	0.250***	-0.028	0.205***	0.130***	0.164**	0.177**	-0.253**
	(0.109)	(0.082)	(0.022)	(0.040)	(0.032)	(0.037)	(0.053)	(0.061)	(0.092)
Median age	-0.116**	-0.066	0.010	-0.071***	-0.007	-0.094***	-0.013	-0.092**	-0.055
	(0.044)	(0.040)	(0.010)	(0.018)	(0.014)	(0.015)	(0.027)	(0.032)	(0.045)
Proportion of male	-0.050	-0.012	0.013	0.019	-0.023	0.009	-0.050*	-0.014	0.078
	(0.042)	(0.033)	(0.009)	(0.016)	(0.013)	(0.014)	(0.024)	(0.027)	(0.041)
Proportion of white	0.017	-0.045	-0.038***	-0.116***	-0.076***	0.003	-0.020	0.004	-0.077
	(0.048)	(0.039)	(0.010)	(0.017)	(0.014)	(0.015)	(0.025)	(0.030)	(0.043)
Intercept	0.313***	0.049	-0.003	0.140***	0.066***	-0.005	0.049	0.298***	0.195***
	(0.050)	(0.046)	(0.011)	(0.021)	(0.017)	(0.018)	(0.030)	(0.040)	(0.049)
Observations	528	650	5949	3309	4863	3473	1788	982	703
R^2	0.193	0.198	0.526	0.081	0.052	0.338	0.040	0.077	0.093
Adjusted R^2	0.177	0.185	0.525	0.078	0.050	0.336	0.035	0.068	0.080
Residual Std. Error	0.882	0.859	0.688	0.894	0.897	0.802	0.954	0.812	0.984
F Statistic	12.366***	15.758***	657.878***	29.159***	26.791***	176.561***	7.462***	8.156***	7.132***

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (two-tailed tests). Excluded the industry *Motion Picture and Video Industries* due to limited number of observations with price and SES data. Excluded *Personal and Laundry Services* due to limited number of observations.

Niche consumption

We use the same process as for New York State to test niche consumption patterns with the data in Texas on brand co-visits for the 1,273 brands. The results are presented in Figure E10. As in New York State, we find no evidence for the niche consumption hypothesis. Apart from few exceptions (e.g., the Cinnabon cluster), the t-SNE plot does not identify distinct consumption niches. Mapping the 10 clusters identified by the k-means algorithm by mean SES and price, we find large consumer clusters for middle and upper-middle SES and more numerous and smaller clusters for low SES. These patterns are the opposite to the expected patterns if niche consumption drives the diverse consumption in high-SES groups.

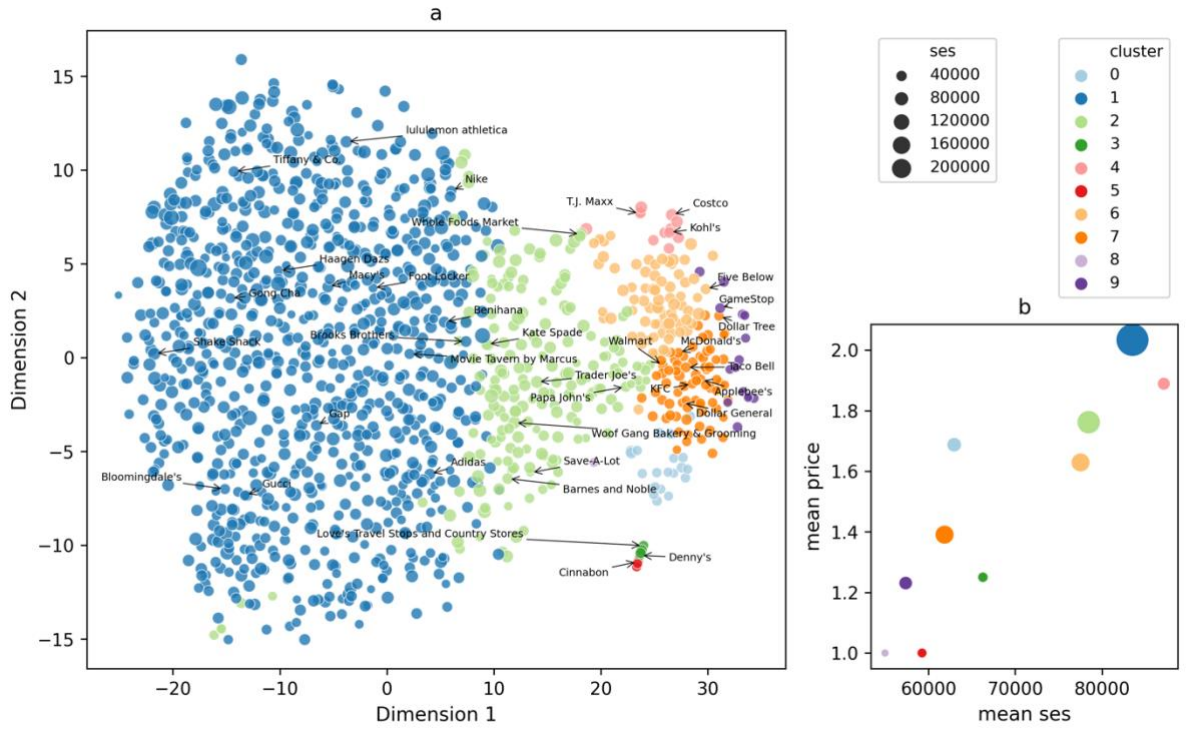

Figure E10. Niche consumption analysis (Texas): a) t-SNE visualization and k-means clustering of the brand co-visit network after node2vec embedding to 128 dimensions; b) mean SES and price for the brands in each cluster.
Note: In panel a, the marker size represents the brand's SES, as the legend indicates. In panel b, the marker size is proportional to the square root of the number of brands in each cluster.

Paper 3 (Chapter 5) Appendices

Appendix A. Matching Results.

Matching Case 1: eight covariates on the full sample

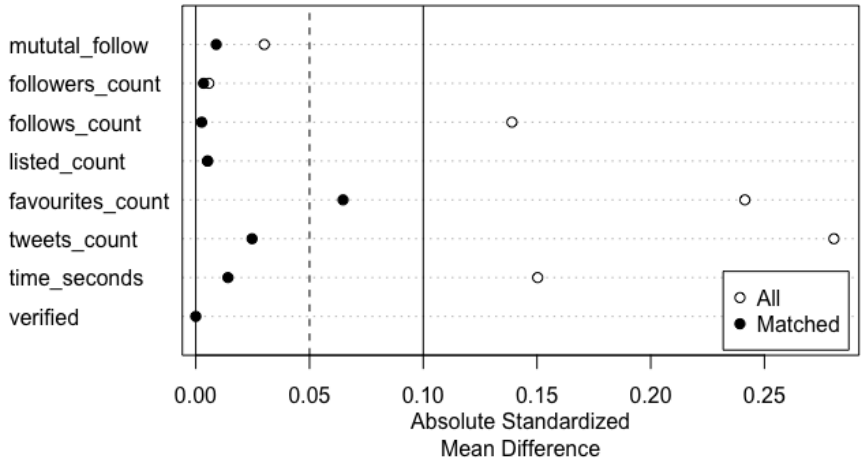

Figure A1. Absolute standardised mean difference of the covariates before and after matching (Case 1).

Table A1. Balance statistics for the matched data (Case 1).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
mututal_follow	-0.0052	-0.0149	0.0090	0.1479	0.0388	0.1699	0.0934
followers_count	0.0010	-0.0023	0.0034	2.2321	0.0355	0.1485	0.0231
follows_count	-0.0156	-0.0184	0.0026	0.1571	0.0226	0.0674	0.1037
listed_count	-0.0058	-0.0110	0.0052	1.5340	0.0050	0.0818	0.0637
favourites_count	0.0258	-0.0368	0.0647	1.2050	0.0596	0.1434	0.1146
tweets_count	-0.0263	-0.0008	-0.0247	0.8752	0.0190	0.0499	0.1148
time_seconds	-0.0085	-0.0225	0.0141	1.0445	0.0097	0.0407	0.1180
verified	0.0263	0.0263	-0.0000	NA	0.0000	0.0000	0.0000

Table A2. Sample sizes (Case 1).

	Control	Treated
All	11606	15839
Matched (ESS)	5977.25	8549.6
Matched	11606	15839
Unmatched	0	0
Discarded	0	0

Matching Case 2: nine covariates on the full sample

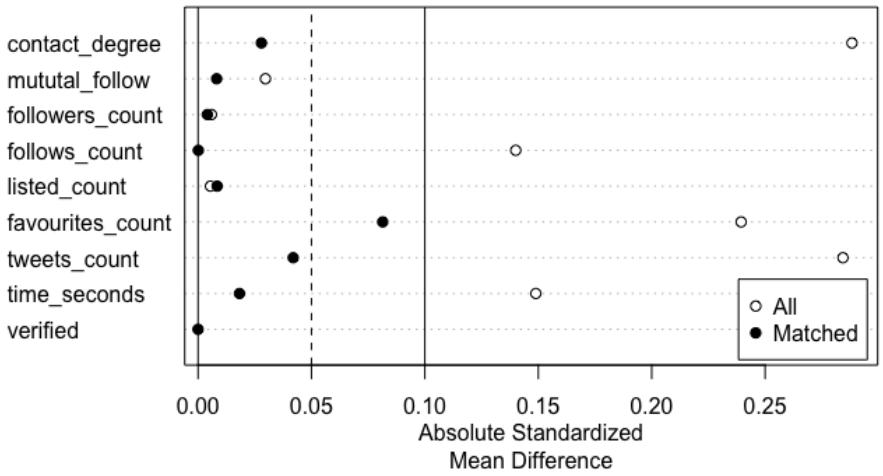

Figure A2. Absolute standardised mean difference of the covariates before and after matching (Case 2).

Table A3. Balance statistics for the matched data (Case 2).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
contact_degree	0.0080	-0.0193	0.0279	1.0713	0.0051	0.0173	0.1545
mututal_follow	-0.0053	-0.0142	0.0082	0.1469	0.0407	0.1892	0.1083
followers_count	0.0012	-0.0028	0.0041	2.2370	0.0383	0.1580	0.0238
follows_count	-0.0181	-0.0182	0.0001	0.1524	0.0229	0.0750	0.1251
listed_count	-0.0044	-0.0128	0.0084	1.5499	0.0061	0.0839	0.0736
favourites_count	0.0277	-0.0506	0.0814	1.2295	0.0703	0.1507	0.1469
tweets_count	-0.0391	0.0041	-0.0419	0.8906	0.0408	0.1028	0.1508
time_seconds	-0.0075	-0.0256	0.0183	1.0499	0.0113	0.0461	0.1551
verified	0.0264	0.0264	-0.0000	NA	0.0000	0.0000	0.0000

Table A4. Sample sizes (Case 2).

	Control	Treated
All	11518	15591
Matched (ESS)	6356.28	9420.54
Matched	11518	15591
Unmatched	0	0
Discarded	0	0

Matching Case 3: nine covariates on the subsample with radius-2 capitals

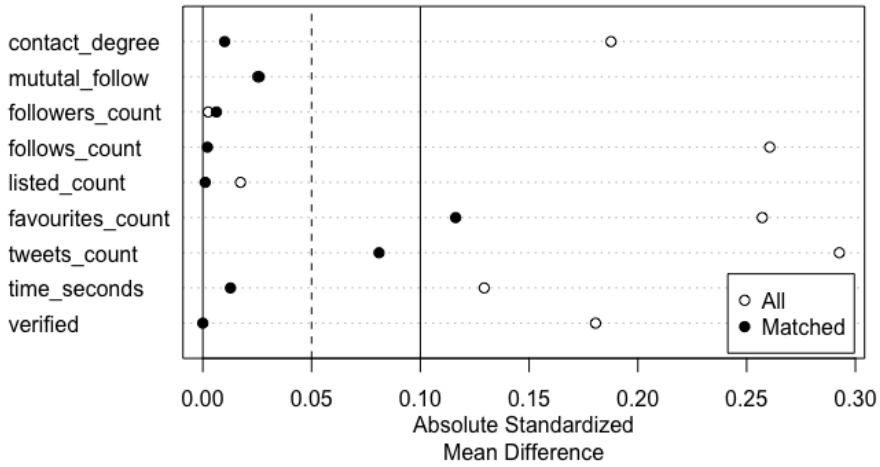

Figure A3. Absolute standardised mean difference of the covariates before and after matching (Case 3).

Table A5. Balance statistics for the matched data (Case 2).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
contact_degree	0.0014	-0.0085	0.0100	1.0482	0.0055	0.0163	0.1981
mututal_follow	0.0050	-0.0217	0.0259	0.4943	0.0584	0.2025	0.1403
followers_count	0.0034	-0.0027	0.0063	2.1987	0.0576	0.1625	0.0270
follows_count	-0.0230	-0.0208	-0.0021	0.4964	0.0243	0.0683	0.1727
listed_count	-0.0063	-0.0052	-0.0011	1.0721	0.0072	0.0797	0.0646
favourites_count	0.0447	-0.0671	0.1162	1.7432	0.1056	0.1857	0.1883
tweets_count	-0.0722	0.0108	-0.0810	0.6925	0.0833	0.1548	0.2061
time_seconds	-0.0108	-0.0234	0.0127	1.0893	0.0147	0.0581	0.2044
verified	0.0179	0.0179	0.0000	NA	0.0000	0.0000	0.0000

Table A6. Sample sizes (Case 3).

	Control	Treated
All	3345	4246
Matched (ESS)	2226.69	2743.4
Matched	3345	4246
Unmatched	0	0
Discarded	0	0

Matching Case 4: eight covariates on the 2022 sample

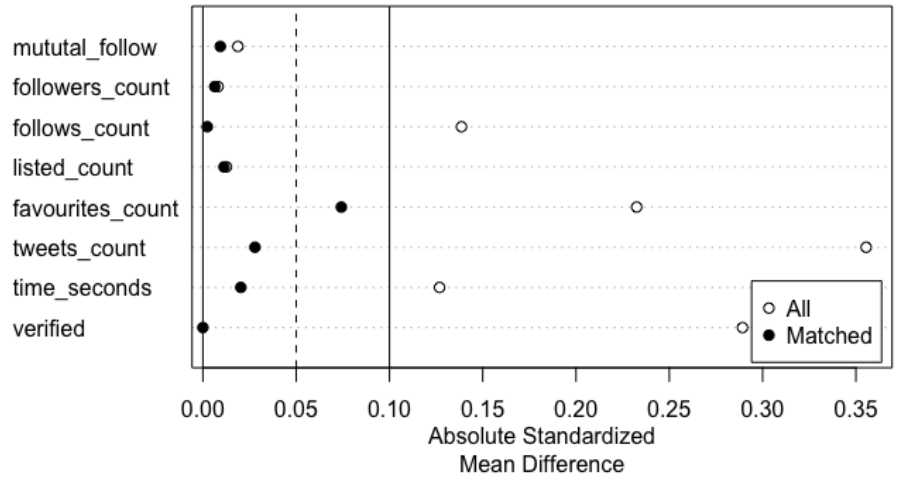

Figure A4. Absolute standardised mean difference of the covariates before and after matching (Case 4).

Table A7. Balance statistics for the matched data (Case 4).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
mututal_follow	-0.0042	-0.0145	0.0093	0.0676	0.0591	0.2226	0.1058
followers_count	0.0008	-0.0051	0.0064	9.3885	0.0544	0.1873	0.0290
follows_count	-0.0188	-0.0162	-0.0023	0.0805	0.0317	0.0978	0.1167
listed_count	-0.0059	-0.0168	0.0113	2.1030	0.0065	0.0937	0.0863
favourites_count	0.0289	-0.0427	0.0742	1.2583	0.0643	0.1361	0.1329
tweets_count	-0.0304	-0.0013	-0.0279	0.8950	0.0211	0.0536	0.1310
time_seconds	-0.0078	-0.0280	0.0203	1.0368	0.0111	0.0452	0.1395
verified	0.0366	0.0366	-0.0000	NA	0.0000	0.0000	0.0000

Table A8. Sample sizes (Case 4).

	Control	Treated
All	7030	10366
Matched (ESS)	3407.3	5687.5
Matched	7030	10366
Unmatched	0	0
Discarded	0	0

Matching Case 5: nine covariates on the 2022 sample

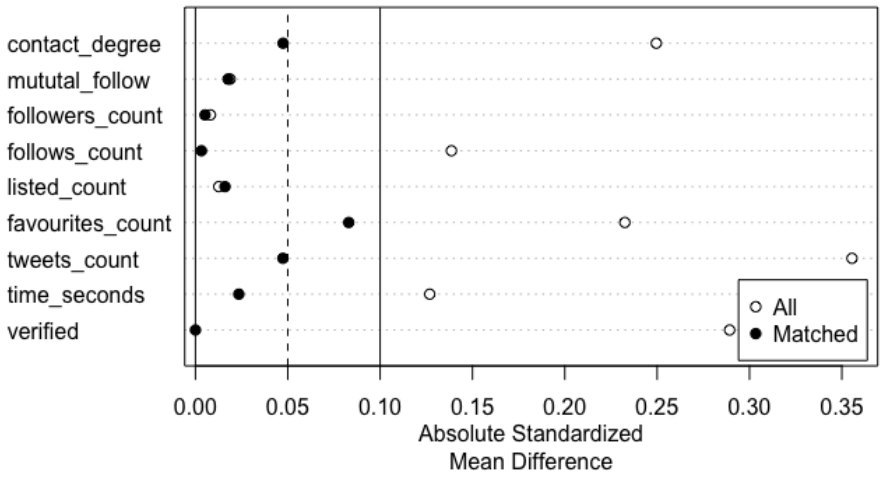

Figure A5. Absolute standardised mean difference of the covariates before and after matching (Case 5).

Table A9. Balance statistics for the matched data (Case 5).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
contact_degree	0.0148	-0.0312	0.0475	1.0827	0.0066	0.0336	0.1641
mututal_follow	-0.0030	-0.0224	0.0177	0.0755	0.0644	0.2425	0.1151
followers_count	0.0007	-0.0041	0.0052	9.3282	0.0582	0.1994	0.0311
follows_count	-0.0201	-0.0237	0.0032	0.0833	0.0347	0.1081	0.1330
listed_count	-0.0063	-0.0218	0.0160	2.1276	0.0080	0.1022	0.0976
favourites_count	0.0316	-0.0485	0.0830	1.2798	0.0696	0.1401	0.1619
tweets_count	-0.0432	0.0062	-0.0474	0.8469	0.0389	0.0810	0.1611
time_seconds	-0.0077	-0.0310	0.0235	1.0719	0.0152	0.0543	0.1764
verified	0.0366	0.0366	-0.0000	NA	0.0000	0.0000	0.0000

Table A10. Sample sizes (Case 5).

	Control	Treated
All	7030	10366
Matched (ESS)	3519.96	5749.26
Matched	7030	10366
Unmatched	0	0
Discarded	0	0

Matching Case 6: nine covariates on the 2022 subsample with radius-2 capitals

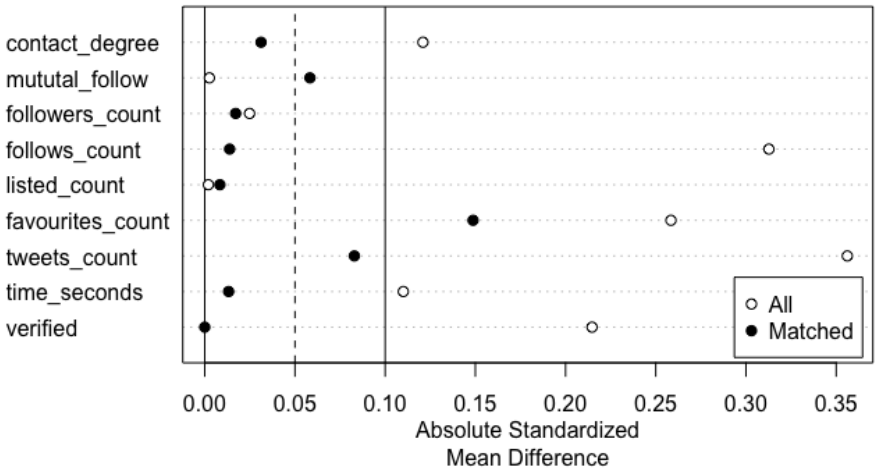

Figure A6. Absolute standardised mean difference of the covariates before and after matching (Case 6).

Table A11. Balance statistics for the matched data (Case 6).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
contact_degree	0.0054	-0.0255	0.0312	1.0725	0.0052	0.0265	0.2062
mututal_follow	0.0193	-0.0414	0.0583	0.5601	0.0858	0.2377	0.1689
followers_count	0.0046	-0.0116	0.0171	12.7066	0.0857	0.2095	0.0324
follows_count	-0.0219	-0.0360	0.0138	0.5548	0.0350	0.0825	0.2077
listed_count	-0.0071	-0.0151	0.0084	4.0185	0.0128	0.1018	0.0865
favourites_count	0.0566	-0.0859	0.1487	1.8715	0.1114	0.1801	0.2319
tweets_count	-0.0772	0.0079	-0.0829	0.6885	0.0720	0.1449	0.2270
time_seconds	-0.0133	-0.0264	0.0132	1.1132	0.0191	0.0706	0.2371
verified	0.0248	0.0248	0.0000	NA	0.0000	0.0000	0.0000

Table A12. Sample sizes (Case 6).

	Control	Treated
All	2078	2756
Matched (ESS)	1380.62	1807.99
Matched	2078	2756
Unmatched	0	0
Discarded	0	0

Appendix B. Correlations between measures.

Table B1. Correlations between social capital measures.

	Degree	Reciprocity	entropy	Local Clustering Coefficient	Effective size	Efficiency
Degree	1	-0.123***	0.153***	-0.033**	0.993***	-0.01
Reciprocity	-0.123***	1	0.289***	-0.09***	-0.128***	0.113***
Entropy	0.153***	0.289***	1	-0.134***	0.08***	0.139***
Local Clustering Coefficient	-0.033**	-0.09***	-0.134***	1	-0.102***	-0.97***
Effective size	0.993***	-0.128***	0.08***	-0.102***	1	0.072***
Efficiency	-0.01	0.113***	0.139***	-0.97***	0.072***	1

Note: *** $p < 0.001$ (two-tailed tests).

Table B1. Correlations between social capital measures constructed with @mentions in 2022.

	Degree	Reciprocity	entropy	Local Clustering Coefficient	Effective size	Efficiency
Degree	1	-0.243***	0.033***	-0.006	0.58***	-0.141***
Reciprocity	-0.243***	1	0.336***	-0.008	-0.332***	-0.058***
Entropy	0.033***	0.336***	1	-0.123***	-0.157***	-0.095***
Local Clustering Coefficient	-0.006	-0.008	-0.123***	1	-0.089***	-0.068***
Effective size	0.58***	-0.332***	-0.157***	-0.089***	1	0.351***
Efficiency	-0.141***	-0.058***	-0.095***	-0.068***	0.351***	1

Note: *** $p < 0.001$ (two-tailed tests).

Table B3. Correlations between language complexity measures.

	Readability	HD-D	MSTTR	MTLD
Readability	1	0.048***	0.065***	0.208***
HD-D	0.048***	1	0.775***	0.687***
MSTTR	0.065***	0.775***	1	0.811***
MTLD	0.208***	0.687***	0.811***	1

Note: *** $p < 0.001$ (two-tailed tests).

Appendix C. Topic modelling details.

The top ten words associated with each topics (probability*words)

'international_affairs'
(0.007*"china" + 0.007*"russia" + 0.006*"ukraine" + 0.005*"russian" + 0.005*"israel" + 0.005*"you.s" + 0.005*"military" + 0.004*"government" + 0.004*"uk" + 0.004*"canada")

'streaming_gaming'

(0.027*"liked" + 0.013*"stream" + 0.008*"twitch" + 0.008*"pc" + 0.006*"xbox" +
0.006*"youtube" + 0.005*"streaming" + 0.004*"gaming" + 0.004*"channel" + 0.003*"ps4")

'covid19'
(0.014*"covid" + 0.012*"vaccine" + 0.011*"patient" + 0.009*"hospital" + 0.008*"medical" +
0.007*"healthcare" + 0.007*"risk" + 0.007*"pandemic" + 0.006*"covid-19" +
0.006*"doctor")

'instantwingame'
(0.244*"instantwingame" + 0.121*"ton" + 0.058*"spinning" + 0.042*"sweep" +
0.037*"freecash" + 0.027*"sweepstakes" + 0.022*"dailyentry" + 0.009*"instant" +
0.009*"instantly" + 0.008*"50.00")

'giveaway_pc_gaming'
(0.110*"giveaway" + 0.041*"entered" + 0.022*"gaming" + 0.013*"pc" +
0.012*"sweepstakes" + 0.010*"international" + 0.009*"contest" + 0.008*"gungiveaway" +
0.008*"rifle" + 0.007*"x")

'tech_industry'
(0.008*"data" + 0.008*"marketing" + 0.005*"technology" + 0.005*"customer" +
0.005*"content" + 0.005*"digital" + 0.004*"industry" + 0.004*"innovation" +
0.004*"startup" + 0.004*"ceo")

'food_promotions',
(0.046*"sponsored" + 0.043*"promotion" + 0.029*"recipe" + 0.011*"delicious" +
0.009*"entry" + 0.008*"avocado" + 0.007*"coupon" + 0.006*"salad" + 0.006*"flavor" +
0.005*"earn")

'giveaway'
(0.159*"giveaway" + 0.041*"entered" + 0.014*"contest" + 0.013*"sweepstakes" +
0.006*"gc" + 0.005*"visa" + 0.005*"mattress" + 0.005*"sweep" + 0.004*"package" +
0.004*"giftcard")

'elections'
(0.007*"election" + 0.005*"county" + 0.005*"voter" + 0.005*"worker" + 0.005*"police" +
0.005*"court" + 0.005*"senate" + 0.004*"justice" + 0.004*"candidate" + 0.004*"congress")

'closet'
(0.090*"added" + 0.064*"poshmark" + 0.062*"shopmycloset" + 0.051*"closet" +
0.043*"wordle" + 0.035*"playlist" + 0.023*"fashion" + 0.023*"item" + 0.015*"4/6" +
0.013*"listing")

'party_politics_emotions'
(0.009*"republican" + 0.007*"gop" + 0.006*"biden" + 0.004*"election" + 0.004*"stupid" +
0.004*"democrat" + 0.003*"idiot" + 0.003*"racist" + 0.003*"putin" + 0.003*"lying")

'diy'
(0.016*"diy" + 0.014*"blog" + 0.009*"disney" + 0.009*"design" + 0.008*"toy" +
0.007*"kitchen" + 0.007*"diaper" + 0.006*"clothdiapers" + 0.006*"craft" +
0.005*"woodworking")

'discourse_markers'
(0.003*"literally" + 0.003*"thread" + 0.003*"weird" + 0.003*"folk" + 0.002*"honestly" + 0.002*"okay" + 0.002*"apparently" + 0.002*"quite" + 0.002*"etc" + 0.002*"completely")

'retweet_ipad_kindle'
(0.134*"rt" + 0.024*"ipad" + 0.023*"follower" + 0.015*"retweet" + 0.013*"swag" + 0.012*"ty" + 0.011*"mt" + 0.010*"buck" + 0.009*"kindle" + 0.009*"+k")

'entertainment'
(0.007*"film" + 0.004*"writing" + 0.004*"writer" + 0.004*"author" + 0.004*"album" + 0.003*"artist" + 0.003*"interview" + 0.002*"david" + 0.002*"novel" + 0.002*"lord")

'strong_emotions'
(0.022*"shit" + 0.018*"fuck" + 0.014*"fucking" + 0.010*"damn" + 0.009*"as" + 0.008*"lt" + 0.007*"lmao" + 0.007*"ya" + 0.007*"dude" + 0.006*"omg")

'education_research'
(0.009*"education" + 0.008*"research" + 0.005*"leadership" + 0.005*"data" + 0.005*"policy" + 0.005*"study" + 0.005*"science" + 0.004*"teaching" + 0.004*"resource" + 0.004*"impact")

'conference'
(0.008*"register" + 0.008*"pm" + 0.007*"award" + 0.005*"conference" + 0.005*"annual" + 0.004*"partner" + 0.004*"thursday" + 0.004*"october" + 0.004*"june" + 0.004*"session")

'sports'
(0.007*"player" + 0.006*"football" + 0.005*"gameinsight" + 0.005*"nascar" + 0.005*"coach" + 0.004*"racing" + 0.004*"driver" + 0.003*"nfl" + 0.003*"baseball" + 0.003*"match")

'sweepstakes'
(0.083*"sweepstakes" + 0.041*"entered" + 0.026*"giveaway" + 0.009*"contest" + 0.009*"entry" + 0.008*"sweep" + 0.007*"ultimate" + 0.007*"1,000" + 0.006*"grand" + 0.006*"10,000")

'party_politics'
(0.005*"political" + 0.004*"government" + 0.004*"policy" + 0.003*"biden" + 0.003*"republican" + 0.003*"election" + 0.002*"evidence" + 0.002*"democracy" + 0.002*"conservative" + 0.002*"argument")

'giveaway_scarf_necklace'
(0.162*"giveaway" + 0.073*"entered" + 0.020*"perduecrew" + 0.007*"amazongiveaway" + 0.005*"scarf" + 0.005*"led" + 0.004*"necklace" + 0.003*"blanket" + 0.003*"coupon" + 0.003*"sunbeltbakery")

'cities'
(0.029*"dc" + 0.020*"ohio" + 0.018*"pa" + 0.015*"virginia" + 0.014*"nj" + 0.013*"philly" + 0.013*"sticker" + 0.012*"va" + 0.011*"baltimore" + 0.010*"mn")

'reward_action'
(0.113*"checked" + 0.097*"download" + 0.072*"earning" + 0.050*"automatically" + 0.049*"mplusplaces" + 0.047*"followed" + 0.043*"mplusrewards" + 0.037*"android" + 0.031*"unfollowed" + 0.028*"reward")

'house_construction'
(0.010*"construction" + 0.008*"customer" + 0.005*"window" + 0.005*"design" + 0.005*"couponing" + 0.004*"safety" + 0.004*"wood" + 0.003*"contractor" + 0.003*"roof" + 0.003*"repair")

'gratitude'
(0.007*"grateful" + 0.006*"quote" + 0.005*"prayer" + 0.005*"joy" + 0.005*"peace" + 0.004*"sending" + 0.004*"appreciate" + 0.004*"lt" + 0.003*"amen" + 0.003*"incredible")

'nyc_ca_chicago'
(0.015*"nyc" + 0.012*"san" + 0.011*"posted" + 0.009*"york" + 0.008*"california" + 0.007*"chicago" + 0.007*"ca" + 0.006*"badge" + 0.006*"bike" + 0.006*"ny")

'climate'
(0.021*"climate" + 0.007*"oil" + 0.006*"gas" + 0.005*"science" + 0.005*"plant" + 0.005*"solar" + 0.004*"fuel" + 0.004*"animal" + 0.004*"farmer" + 0.004*"environmental")

'happy_houselife'
(0.003*"snow" + 0.003*"lovely" + 0.002*"omg" + 0.002*"husband" + 0.002*"chocolate" + 0.002*"yay" + 0.002*"cheese" + 0.002*"fantastic" + 0.002*"cream" + 0.002*"cake")

'sweepstakes_a1_a5'
(0.020*"sweepstakes" + 0.012*"finger" + 0.011*"fancaveentry" + 0.010*"a2" + 0.010*"a1" + 0.010*"a3" + 0.009*"a4" + 0.009*"crossed" + 0.009*"a5" + 0.009*"contest")

Merged topic groups

'merged_promotions': ['instantwingame', 'giveaway_pc_gaming', 'food_promotions', 'giveaway', 'retweet_ipad_kindle', 'retweet_ipad_kindle', 'sweepstakes', 'giveaway_scarf_necklace', 'reward_action', 'sweepstakes_a1_a5']

'merged_politics': ['elections', 'international_affairs', 'party_politics', 'party_politics_emotions']

'merged_cities':['nyc_ca_chicago', 'cities'],

'merged_entertainment': ['entertainment', 'sports', 'streaming_gaming']

Appendix D. Words and hashtags about immigration.

'immigrant', 'immigration', 'refugee', 'foreigner', 'asylum seeker', 'undocumented worker', 'foreign worker', 'illegal alien', 'illegal worker', 'asylumseeker', 'undocumentedworker', 'foreignworker', 'illegalalien', 'illegalworker', 'leavenoonebehind', 'nowall','noborders',

'openborders', 'familiesbelongtogether', '#illegals', 'keepthemout', 'ourcountry', 'sendthemback', 'theyhavetogoback', 'deportthemall'

Appendix E. Matching results for the immigration and baseline sample.

The immigration sample

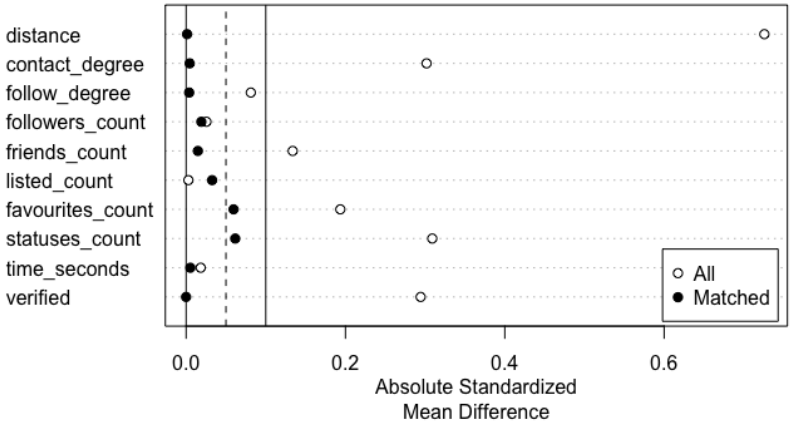

Figure E1. Absolute standardised mean difference of the covariates before and after matching (the immigration sample).

Table E1. Balance statistics for the matched data (the immigration sample).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.8585	0.8583	0.0012	0.9816	0.0008	0.0130	0.0040
contact_degree	0.0133	0.0176	-0.0045	1.0029	0.0050	0.0195	0.8495
mututal_follow	-0.0101	-0.0032	-0.0038	0.1495	0.0631	0.2370	0.1150
followers_count	0.0002	-0.0144	0.0191	301.7070	0.0729	0.2132	0.0479
follows_count	-0.0108	-0.0375	0.0148	0.1229	0.0692	0.2017	0.1438
listed_count	-0.0014	-0.0307	0.0326	6.9144	0.0160	0.1172	0.1588
favourites_count	0.0191	-0.0338	0.0596	1.1038	0.0718	0.1268	0.4708
tweets_count	0.0301	0.1088	-0.0616	1.1716	0.0844	0.1440	0.4430
time_seconds	-0.0030	0.0023	-0.0054	1.0840	0.0127	0.0317	1.0972
verified	0.0485	0.0485	-0.0000	NA	0.0000	0.0000	0.0000

Table E2. Sample sizes (the immigration sample).

	Control	Treated
All	1457	8827
Matched (ESS)	591.94	8082.28
Matched	1457	8827
Unmatched	0	0
Discarded	0	0

The baseline sample

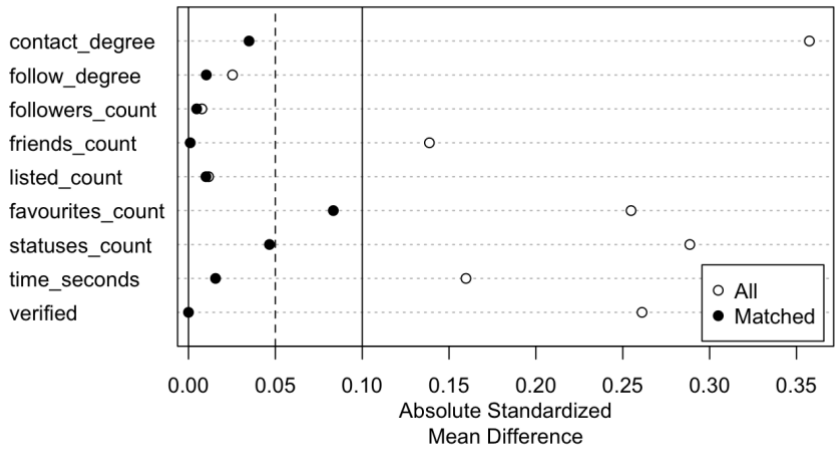

Figure E2. Absolute standardised mean difference of the covariates before and after matching (the baseline sample).

Table E3. Balance statistics for the matched data (the baseline sample).

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
contact_degree	0.0106	-0.0233	0.0348	1.0726	0.0061	0.0188	0.1594
mututal_follow	-0.0060	-0.0169	0.0103	0.1442	0.0455	0.2022	0.1096
followers_count	0.0011	-0.0035	0.0047	2.2351	0.0415	0.1729	0.0231
follows_count	-0.0206	-0.0196	-0.0010	0.1457	0.0244	0.0799	0.1288
listed_count	-0.0048	-0.0147	0.0101	1.5485	0.0064	0.0887	0.0744
favourites_count	0.0338	-0.0474	0.0834	1.2142	0.0744	0.1547	0.1489
tweets_count	-0.0418	0.0059	-0.0466	0.8624	0.0427	0.0977	0.1548
time_seconds	-0.0085	-0.0239	0.0156	1.0530	0.0115	0.0481	0.1586
verified	0.0283	0.0283	0.0000	NA	0.0000	0.0000	0.0000

Table E4. Sample sizes (the baseline sample).

	Control	Treated
All	10875	13715
Matched (ESS)	5921.04	7854.89
Matched	10875	13715
Unmatched	0	0
Discarded	0	0

Appendix F. Promotional words and hashtags.

Promotional words: 'giveaway', 'sweepstake', 'enter to win', 'sweep', 'entering to win', 'chance to win', 'you can win', 'entered to win', 'enter daily', 'giving away'.

Promotional hastags: '#giveaway, #win, #sweepstakes, #instantwingame, #ad, #contest, #promotion, #sponsored, #perduecrew, #sweeps, #entry, #free, #amazongiveaway, #giveaways, #amazon, #giftcard, #shopmycloset, #avosweepstakes, #gungiveaway, #competition, #gggentry, #mplusrewards, #mplusplaces, #gameinsight, #android, #sunbeltbakery, #entertowin, #guacfrommexico, #step2, #christmas, #fishbowlprizes, #mattress, #fancaveentry, #dailyentry, #afmsweepstakes, #countdowntochristmas, #notsorry, #deals, #attexploresweeps, #gotitfree, #royaldraw, #prize, #prizes, #fanfriday, #glamcrowd, #bakeryfreshfriends, #venmome, #androidgames, #poshmark, #winagun, #giftguide, #mpoints, #promo, #enter, #guacworld\t, #trynatural, #700inone, #vivaveltoro, #freebies, #guacit, #splendasavvies, #ipadgames, #listia, #freestuff, #contestalert, #instantwin, #cashappday, #bestof2021sweepstakes, #giveawayentry, #thriftyniftymom, #aaronreckgiveaway, #canwin, #guacthetailgate, #giveawayalert, #doritosroulette, #koasweepstakes, #freesample, #twitterparty, #freesamp, #sweetairchannels, #kohlscashsweepstakes, #attexplore, #giveawayhop, #77inone, #guacamoments, #countdownentry, #guacfood, #5gsfor5g, #giveawayalert, #frankssweepstakes, #spon, #tumsbingosweepstakes, #fastadvilfanatics, #echalevidasweepstakes, #intelrigchallenge, #guncontest, #stpromo, #sweepstakesentry, #crunchclassicentry, #tweets4toys, #moneygrammonday, #sweepsentry, #bestbuytechzonesweepstakes, #bcsweepstakes, #perfectbar, #teamsaturdayentry, #esurancesweepstakes, #deal#winitwednesday, #winit, #givingtuesday, #cashappfriday, #freebie, #avoeatery, #owsentry, #freelitterrobot, #magiveaway, #12daysofgiveaways, #sale, #freeskier, #ilovetmobile, #freebiefriday, #winandyourein, #tweets4toys, #safediggingmonth, #pinchmas, #xpressotours, #viziofans, #enjoymore, #maglite, #rocketmortgagesquares, #sweep, #skinitmade, #smarthome, #enjoymore, #litterrobot, #tmobiletuesdays, #petgiveaway, #keurig, #intelgamerdays, #magtac, #watermelon, #maglitenation, #balsamhill, #scooplesssummer, #greengiantfresh\t, #doggiveaway, #melonmania, #beautyofcarpet, #zellesweepstakes, #rocketmortgagesquares, #greengiantfresh, #samplesource, #ultimateroadtrip, #xboxonex, #playstation5, #greengiantfresh, #generationgood, #madeinusa, #therawin, #oculusquest2, #samplesource, #makeitpop, #19crimes, #kawaiigiveaway, #fishinggiveaway, #xboxone, #kohlsblackfridaysweepstakes, #gotouring, #alienwaresweepstakes, #candygiveaway, #attsweepstakes, #justoneclick, #cashapp13plus, #nowthatsabigdeal, #etrike, #ebikes, #tricycle, #libertytrike, #linksawakening, #bicycle, #25daysofgiving, #iheartradio, #cutthecordday, #4aklondike\t, #mypromotionentry, #mtndewmajormelon, #homefortheholidays, #gascash, #thehealthypotatocompany, #cashappextracredit, #dominosquikly, #whyichime, #12daysofshespeaks, #cashappgifting, #cashappbitcoinchallenge, #cashappboostweek, #coupon, #smiley360, #fortune4days'.

Appendix G. Hypotheses testing results using the subsample without promotional tweets and users.

Table G1. Average differences after matching between high and low SES users for social capital measures.

	Estimate	Standard Error	P	2.50%	97.50%

Degree	6.72	0.654	<0.001	5.44	8
Reciprocity	0.0439	0.00358	<0.001	0.0369	0.0509
Topological diversity	0.0327	0.00224	<0.001	0.0284	0.0371
Local Clustering Coefficient	-0.0104	0.00267	<0.001	-0.0156	-0.00514
Effective Size	0.479	0.0872	<0.001	0.308	0.65
Efficiency	0.0104	0.00226	<0.001	0.00594	0.0148

Table G2. Average differences after matching between high and low SES users for social capital measures constructed with @mentions in 2022.

	Estimate	**Standard Error**	**P**	**2.50%**	**97.50%**
Degree	6.14	0.578	<0.001	5	7.27
Reciprocity	0.0393	0.00546	<0.001	0.0286	0.05
Topological diversity	0.0436	0.00313	<0.001	0.0375	0.0497
Local Clustering Coefficient	-2.21e-07	3.41e-08	<0.001	-2.88e-07	-1.55e-07
Effective Size	26.2	3.71	<0.001	18.9	33.5
Efficiency	4.6	1.1	<0.001	2.44	6.76

Table G3. Average differences after matching between high and low SES users for communication pattern measures.

	Estimate	**Standard Error**	**P**	**2.50%**	**97.50%**
Readability	1.38	0.0297	<0.001	1.32	1.44
HD-D (lexical diversity)	0.0159	0.000784	<0.001	0.0144	0.0174
MSTTR (lexical diversity)	0.0163	0.000609	<0.001	0.0151	0.0175
MTLD (lexical diversity)	38.2	0.775	<0.001	36.7	39.7
Forward Looking	0.003	0.00056	<0.001	0.0019	0.00409