Yuanmo He

+44 7928549255 l y.he54@lse.ac.uk l yuanmohe.com l London

EDUCATION

PhD Social Research Methods (specialising in computational social science) 09/2020 - 09/2024
The London School of Economics and Political Science (LSE)

- Applying advanced data analysis and computational methods (e.g., machine learning, natural language processing, social network analysis) on digital trace data to study social networks, culture, and inequality.
- Supervisors: Dr Milena Tsvetkova and Professor Kenneth Benoit.
- Affiliations: Data Science Institute, International Inequality Institute.
- Awarded the **LSE PhD Studentship** for four years.

MSc Applied Social Data Science (Distinction) 09/2019 - 08/2020
The London School of Economics and Political Science

- Relevant modules: Computer Programming, Data for Data Scientist, Applied Machine learning, Quantitative Text Analysis, Multivariate Analysis and Measurement, Fundamentals of Social Science Research Design.
- **Distinction in all modules**.

BSc Social Sciences (First Class Honours) 09/2016 - 06/2019
University College London (UCL)

- Relevant modules: Social Network Analysis, Causal Analysis in Data Science, Quantitative Research Methods, Cognitive Psychology, Social Psychology, Game Theory.
- Awarded the UCL Institute of Education **Faculty Medal** (the best final year undergraduate student).
- Achieved **the highest final mark** in the Department of Social Science.

Coursera:

- Statistics with R Specialization: probability, inferential statistics, Bayesian statistics 06 - 08/2018
- Mathematics for Machine Learning Specialization: linear algebra, multivariate calculus, PCA 06 - 08/2018

WORKING PAPER

He, Y and Tsvetkova, M. *Estimating Individual Socioeconomic Status of Twitter Users.* (Manuscript available upon request.)

- Based on classical social theories, developed a method that uses correspondence analysis to estimate Twitter users' socioeconomic status based on the brands they follow.
- Worked on a **complete data science workflow**: from data collection, data cleaning, exploratory analysis, model building, results evaluation, to oral and written communication.
- Used R, Python, SQL, Azure Clouding Computing, Twitter API, and Google Geocoding API to collect, process, clean and select **190 million** rows of data and estimated the socioeconomic status of **3,482,657** Twitter users and **339** brands.
- Validated the estimates with data on audience composition from the Facebook Marketing API, self-reported job titles on users' Twitter profiles, and a small sample of survey data. Our measure of socioeconomic status achieved **significant correlation (0.5-0.7)** with income, education, and occupational social class at the aggregated level.

CONFERENCE PRESENTATIONS

Estimating Individual Socioeconomic Status of Twitter Users

- General Online Research, online 09/2021
- The Annual Meeting of the American Sociological Association (Section on Inequality, Poverty and Mobility: New Approaches to Understanding and Addressing Inequality), online 08/2021
- International Conference on Computational Social Science, online 07/2021

TEACHING EXPERIENCE

MY474 Applied Machine Learning, Teaching Assistant, LSE 01/2022 - 04/2022
MY470 Computer Programming, Teaching Assistant, LSE 09/2021 - 01/2022
Introduction to Python Programming, Teaching Assistant, Data Science Summer School 08/2021

PROJECTS

Bayesian Estimation for the Socioeconomic Status of Twitter Users 04/2020 - 08/2020

- Built a latent space model that represents the following network of Twitter users and brands, where the distance between a user and brand depends on their proximity of socioeconomic status. Applied No-U-Turn sampler and Metropolis-Hasting algorithm with R and Stan to estimate the parameters for a network of 360,000 users and 359 brands.

The Social Contagion of Cheating 01/2020

- Created network simulations with Python based on 6,000 match records from the massive multiplayer online game PlayerUnknown's Battleground to test whether the victims of cheater are more likely to cheat.

SKILLS

Programming language & statistical software: Python, R, SQL, Stata*, Stan*, SPSS*

Python packages: NumPy, pandas, scikit-learn (non-exhaustive)

R packages: tidyverse, tm, quanteda, glmnet, randomForest, e1071 (non-exhaustive)

Advanced Data Analysis: machine learning, natural language processing, social network analysis, multivariate analysis, causal analysis, multilevel modelling*, parallel computing*, cloud computing* (**indicates basic skill-level*)

Languages: Chinese (native), English (full professional proficiency)