

International Journal of Population Data Science

Journal Website: www.ijpds.org



Augmenting Surveys with Social Media Data: A Probabilistic Framework for LinkedIn Data Linkage.

Paulo Matos Serodio¹, Tarek Al Baghal¹, Luke Sloan², Shujun Liu², and Curtis Jessop³

¹ISER, University of Essex

²University of Cardiff

³NatCen

Introduction & Background

LinkedIn, with its extensive global network of over 900 million members across more than 200 countries, presents a unique repository for examining labour market dynamics, professional development, and the impact of social networking on employment opportunities. Despite its potential, LinkedIn's wealth of data on professional trajectories, skills, and labour market outcomes remains largely untapped in survey research due to challenges in data collection.

Objectives & Approach

This paper introduces a novel methodology for integrating LinkedIn data with survey responses using data from the fourteenth wave of the Innovation Panel (IP14) of Understanding Society: The UK Household Longitudinal Study (UKHLS), conducted in 2021. In IP14, we probed the extent of LinkedIn usage among the UK population and assessed users' willingness to link their LinkedIn profiles with their survey responses. Those consenting to link their accounts were asked for specific details — namely their first and last names, employer, and job title — to enable profile identification on LinkedIn. Faced with the unavailability of a unique platform identifier and the cessation of LinkedIn's API, this information was crucial for matching profiles accurately.

We crafted a framework using PhantomBuster for ethical data extraction and a probabilistic string-matching technique to ensure precise linkage between survey responses and LinkedIn profiles. PhantomBuster, a cloud-based tool, efficiently scrapes dynamic content using JavaScript in a headless browser environment, sidestepping IP-related restrictions while adhering to website terms of service. It streamlines the data collection process. Identified profiles were subjected to an iterative probabilistic string matching, using respondent-provided metadata alongside supplementary data, to maximize the accuracy of matching the profiles to our survey participants.

Relevance to Digital Footprints

The described method advances digital footprint research in data collection and linkage. It automates the retrieval of vast online data sets; compiles information efficiently in an organized format; saves time and labour by mechanizing monotonous tasks; circumvents platform-imposed IP restrictions; and imposes fewer barriers to entry as it requires less technical skill than other scraping tools like Selenium.

Conclusions & Implications

This approach not only facilitates the precise identification and collection of LinkedIn profile data but also sets a precedent for ethical considerations in web scraping practices. By documenting this methodology, we aim to equip researchers with a scalable and replicable tool for future studies, enriching the analysis of labour market outcomes and the interplay between formal education, informal training, and professional success through the integration of LinkedIn and survey data.

