GRAPH-EE: BUILDING EMPLOYER - EMPLOYEE PANELS FROM A KNOWLEDGE GRAPH AND COMPANY MICRODATA

Abstract

Linked employer-employee datasets are a hugely powerful tool for researchers and policymakers across the social sciences. However, access to such data is typically highly restricted and the quality of data varies substantially across countries. This paper presents GRAPH-EE, a large-scale employer-employee platform for larger companies in the UK. Leveraging a vast knowledge graph of the global public internet linked to UK company microdata and global patents data, GRAPH-EE is a rich complement to administrative data sources. This version of the platform covers over 10,500 companies and 800,000 workers active in the UK during 2007-2023. We show the potential to extend our approach to multiple other countries. A public release version of the data, plus codebase and documentation, is available to researchers.

Keywords: data science, employer-employee data, productivity, innovation, skills

JEL codes: C55, C81, J10, J24, J61, L25, O31

Author info

Steve Gray, UCL
Tom Kemeny, University of Toronto
Max Nathan, UCL and CEP. Corresponding author. max.nathan@ucl.ac.uk
Ceren Ozgen, University of Birmingham
Guido Pialli, University of Turin
Jon Reades, UCL
Anna Rosso, University of Milan
Mateo Seré, UCL
Sameera Siddiqui, UCL
Anna Valero, LSE and CEP

This version: September 2026. Click here for latest version

Acknowledgements

Thanks to audiences at GEOINNO 2024 and 2026, GCEG 2025, RSA Winter 2025, and workshops at UKRI, UCL and the Universities of Birmingham, Hamburg, Reading and Ca'Foscari Venice for helpful comments; to Rebecca Lee at OpenCorporates and Filipe Mesquita at Diffbot for extensive advice; and to our project advisory board for constructive feedback. This project uses data from Diffbot, OpenCorporates, Orbis Historical, Orbis IP and PATSTAT Global. Thanks to all data providers. This research is funded by UKRI Grant ES/W010232/1, Diversity and UK Firm Performance. The project was reviewed and approved by the UCL Research Ethics Board, application 22883/00. This paper represents the views of the authors, not the advisers, data providers or funders.

Author contributions

Steve Gray: data curation, resources, software; Max Nathan: conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, writing - original draft; Tom Kemeny: conceptualisation, methodology, writing - reviewing and editing; Ceren Ozgen: conceptualisation, methodology; Guido Pialli: data curation, formal analysis, methodology, software, validation, visualisation, writing - original draft; Jon Reades: conceptualisation, data curation, resources, software; Anna Rosso: conceptualisation, data curation, formal analysis, methodology, software, writing - reviewing and editing; Mateo Seré: data curation, formal analysis, methodology, software, validation, visualisation, writing – original draft; Sameera Siddiqui: data curation, validation; Anna Valero: conceptualisation, methodology, writing - reviewing and editing. Authors are listed alphabetically.

Competing interests

The authors declare no competing interests.

1. Background & Summary

Linked employer-employee datasets are a hugely powerful tool for researchers and policymakers. They allow us to track workers through their careers and across locations, and to look at the co-evolution of firms' workforces with other characteristics and outcomes. In turn, this opens up a range of questions that cannot be easily answered without data of this kind. Specifically, researchers can investigate workforce drivers of firm productivity, profits and sales; earnings and human capital accumulation; career progression and the role of gender, qualifications and family background in explaining this; labour flows across places, including flows of workers into and out of cities, and the economic returns to working in different locations; and the characteristics and drivers of knowledge-intensive clusters (Card et al., 2025; De La Roca & Puga, 2017; Goldin et al., 2017; Haltiwanger et al., 1999; Heyman et al., 2007; Kemeny & Cooke, 2018; Kemeny et al., 2020). When linked to other datasets, such as school records, patents data and firm / workplace surveys, worker-firm spines become a form of 'administrative big data', allowing even richer insights.[1]

In practice there are two main constraints on what we can learn from such data. First, these data typically contain confidential and personal sensitive information, are held securely and have stringent conditions limiting access and use. Firms and workers are anonymized, linkage to other data may be forbidden or restricted, and only non-disclosive results can be reported. Second, the size and richness of the data vary considerably in practice. Some countries provide complete registers of firms and workers; others provide much smaller surveys which may only cover a few thousand firms and their workforces. See Table 1 for selected examples. The set of firm and worker characteristics also varies widely; some countries provide very rich vectors of company and worker data; others provide very limited information. In turn, this creates binding constraints on cross-national analysis, which has very high co-ordination costs: 'distributed microdata' designs develop a high-level common sampling frame and workflow, which is then implemented by teams of

[1] See for example Longitudinal Education Outcomes database and the linked datasets produced by the Wage and Employment Dynamics project (https://www.gov.uk/government/publications/longitudinal-education-outcomes-leo-dataset/longitudinal-education-outcomes-leo-data; https://www.wagedynamics.com). Both accessed 8 December 2025.

researchers in each study country to generate broadly harmonised results (Barreto et al., 2025; Boustan et al., 2025; Criscuolo et al., 2021).

In the UK, for example, researchers currently have access to two employer-employee sources. The Annual Survey of Hours and Earnings (ASHE) covers around 180,000 workers, providing rich information on pay and hours but very limited information on worker or firm characteristics (Office for National Statistics, 2025a). The Longitudinal Educational Outcomes (LEO2) dataset is a complete register of education and labour market history, which can be linked to a small set of firm-level outcomes`, but is limited to those born after 1985 lacks information on occupation and does not link work spells to employer (Department for Education et al., 2025). The UK Office of National Statistics plans to release a full worker-firm 'spine' building on LEO2, but on a gradual and restricted basis and no confirmed timeline. These constraints are also visible in countries with better basic provision than the UK. For example, the US Longitudinal Employer Household Dynamics (LEHD) panel only covers certain states over particular periods, provides very limited information on worker and establishment characteristics, and descriptive results can only be shown for highly aggregated, multi-state geographies(Vilhuber, 2018). The Norwegian LEED panel provides very rich information but for a small, homogenous country, and is limited on individual characteristics, including education and qualifications.[2] Dutch administrative data covers the universe of firms and workers over a 20-year period, but has limited coverage of occupation and educational information.[3]

Many of these issues are structural and cannot be straightforwardly 'solved'. In this scenario, the combination of online and open administrative sources can provide a rich complement to conventional worker-firm datasets. The core workflow involves one or more of a) building individual characteristics, education and career histories from web data, b) linking workers to companies, and c) enriching the company-level data using open and/or commercial sources. A growing literature is already developing along these lines, including both proof-of-

[2] The Norwegian Matched Employer-Employee Register is based on State Register of Employers and Employees (the Aa Register), maintained by the Norwegian Tax Administration (NAV), and augmented with data from Statistics Norway.

[3] https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research/overview-of-all-datasets

concept studies and quantitative analysis on a range of firm and worker outcomes (Amazadeh et al., 2024; Babina et al., 2022; Breithaupt et al., 2025; Dahlke et al., 2025; Dorn et al., 2025; Fedyk & Hodson, 2022; Gagliardi et al., 2024; Jeffers, 2024; Jin et al., 2025; Lee & Glennon, 2023; Rock, 2019; Tambe, 2014). The majority of these studies cover the US, leverage commercial data on online worker resumes (such as Revelio, Cognism and LinkedIn) and sometimes match this to large companies in commercial databases like Compustat or Orbis (Babina et al., 2022; Dorn et al., 2025; Fedyk & Hodson, 2022; Gagliardi et al., 2024; Jeffers, 2024; Rock, 2019; Tambe, 2014). Linkages across these data sources represents a major practical challenge in this emerging literature, including matching individuals to companies, and linking company microdata to company web profiles. Studies typically need to deploy elaborate fuzzy matching routines to achieve these connections, with high noise and attrition rates. This limits both the size of the resulting data, and what can be learnt from it.

We combine information from a world-leading knowledge graph – essentially, a database storing linked entities and their relationships – with open company data to build GRAPH-EE, a unique employer-employee data platform for the UK. We draw rich worker and organisation data from Diffbot, a commercial knowledge graph of the public internet. We exploit the fact that the UK companies register is open data, with company identifiers in Diffbot's graph. We use this to precisely link companies and workers in Diffbot to companies in the UK register, to rich company financial data from Orbis Historical and to patents data from PATSTAT Global and Orbis IP. The resulting core data comprises over 10,500 larger UK companies active between 2007 and 2023, and over 800,000 workers in these companies. Specifically, because we need information on company headcounts, we focus on UK companies which are legally required to provide full, audited accounts (see Section 2).

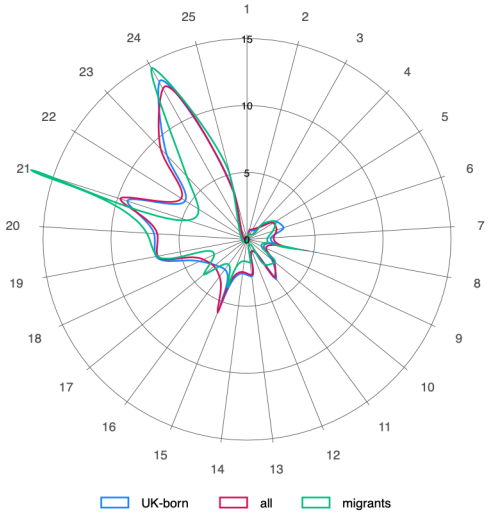
GRAPH-EE can provide very rich insights. For example, Figure 1 summarises multidimensional skills for over 650,000 workers in our sample firms, using topic modelling to organize 32,000 individual skills strings into 25 topics and using LLMs to classify topics according to skill complexity, as represented in job descriptors in the ISCO occupational classification. Section 3 provides more details of the build. In Section 4 we show how these measures of workforce skills are distinct from measures of formal human capital typically measured in

administrative data, and correlate with years of labour market experience, consistent with theories of human capital development (Acemoglu & Autor, 2011; Autor, 2013; Becker, 1962). Panel A shows wordclouds for each topic. The skills space runs from general skills for office workers (e.g. rows 1 and 2), through managerial skills (e.g. rows 3 and 4) to specialized skills (row 5). Panel B shows the distribution of topics across all workers (red line) and compares migrants against UK-born workers (green vs. blue lines). We can see that migrant workers have larger shares of both the most complex and least complex skills.

Figure 1. Text-based worker skills: topics and distribution across worker types.

A. Wordclouds for skills topics

B. Topic shares by worker type



Source: Diffbot, authors. Source data is 32,000 skills strings for ~ 655,000 workers observed in larger UK firms, 2007-2023. Each worker is observed once. In Panel A, each wordcloud represents one of 25 topics generated through LDA analysis of skills strings. *N*-grams represent the most frequent terms in each topic. Topics are labelled and ordered by role complexity using GPT-4o, where Topic 1 is the least complex (top left) and Topic 25 is the most complex (bottom right). Panel B is a spider graph showing the distribution of these topics across worker types. For each worker, we assign a 'dominant topic' as the one they are most likely to have, based on observed Diffbot skills. Spikes denote topics 1-25, where 1 is the least complex and 25 is the most complex. Grid rings denote percentages 1-100.

Table 1. Employer-employee and open company data: selected OECD countries

Country	Employer-employee administrative data	Company register in OpenCorporates	Company identifiers in Diffbot
Canada	Register	Yes, open	
Denmark	Register	Yes, open	
Finland	Register	Yes, open	
France	Register	Yes, open	SIRENE identifiers
Germany	Survey	Yes	
Ireland	Survey	Yes, open	
Israel	Yes	Yes, open	Company numbers
Italy	Yes	Yes	
Japan	Partial register	Yes, open	Corporate numbers
Netherlands	Register	Yes	
Singapore	None	Yes, open	Unique entity number
South Korea	Survey	Yes	
Spain	Survey	No	
UK	Partial register, survey	Yes, open	Company Reference numbers
USA	Partial register	Yes, varies by state	IRS identifiers

Source: Authors' elaboration, OpenCorporates, Diffbot. Notes: 1) 'Employer-employee administrative data' based on Google Scholar search: 'employer-employee data + [country name]', then hand-coded based on reading data section of articles returned on first page of search results. 2) 'Company register in OpenCorporates' based on http://registries.opencorporates.com, a leading aggregator of global company register data. Marked as 'yes' if all companies are included and are freely available. Marked as 'yes, open' if data is freely downloadable / accessible through an API. Otherwise, data is available but access requires permissions and/or cost. 3) 'Company identifiers in Diffbot' based on https://docs.diffbot.com/docs/ont-organization metadata. Marked if currently included in Diffbot. Otherwise empty.

Table 1 gives a sense of the future potential of our approach, for selected OECD countries. Most provide administrative employer-employee data, but there is wide variation in terms of the size and richness of these data. All countries provide company register data, and often this is open, allowing researchers to easily extract and work with it along the lines above. Only a minority of these registers currently exist in the Diffbot knowledge graph, our core data source, but all the open registers could be incorporated straightforwardly, and others could be included given provider permissions. Overall, our workflow could be used to provide rich and flexible complements to conventional worker-firm data in almost all of these cases.

The paper is structured as follows. Section 2 describes our key data sources. Section 3 sets out the build. Section 4 summarises the key variables in the full and public builds. Section 5 describes validation exercises. Section 6 concludes. A public release version of GRAPH-EE, plus codebase, documentation and selected input layers, is available via the UK Data Service ReShare repository, collection ID 858769.

2. Data

To build the platform we combine a) company and worker profiles and histories from the graph with b) UK company microdata and c) patents microdata. Section 2.1 describes our data sources. Section 2.2 sets out our workflow.

2.1 Diffbot

We use Diffbot to build company profiles, worker profiles and worker education and career histories, through paid-for access to Diffbot's interfaces and APIs. Diffbot is a commercial knowledge graph database of the entire public web. For firms and workers, it will typically draw on public information on open company registries, company websites (e.g. company and worker profiles), business intelligence websites, business directories, public social media profiles, and media coverage of firms and individuals.[4] At the start of 2025, the graph included 278.9m active

[4] https://www.diffbot.com/products/knowledge-graph/.

companies and 231.3m individuals in employment worldwide.[5] In the UK, the graph covered 4.7m active companies with company identifiers and 10.6m workers.[6] This compares to totals of 5.5m active firms in the UK Business Population Estimates, and 33.9m workers aged 16+ in the UK Labour Force Survey (Department for Business and Trade, 2025; Office for National Statistics, 2025b). In Section 3 we run further diagnostics, benchmarking Diffbot against administrative data.

Knowledge graphs were an established tool in early AI research, and a key concept in early visions of the semantic web. Developments in machine learning mean that automated graphs are now feasible: Google routinely uses knowledge graphs to present popular search results; academic studies have used LLM-generated graphs to map value chains (Douglas Heaven, 2020; Fetzer et al., 2024). Diffbot builds its graph by continually crawling the public web, identifying key elements on webpages, then using image recognition, natural language processing and supervised learning to build a graph of entities (such as people, organisations, places), their characteristics and relationships to each other (Mesquita et al., 2019). Specifically, Diffbot builds proprietary tools based on knowledge fusion algorithms, which use supervised learning to infer properties and linkages from high-quality sources in previous versions of the graph, ranking items using a confidence score (Dong et al., 2014).

In our case, Diffbot allows us to track a company and its workforce over time, as well as seeing an array of individual and firm-level characteristics. We exploit three features of Diffbot in particular. First, as the UK Company Register is provided as open data, the graph includes detailed UK company information, including Company Registration Number (CRN) identifiers. These identifiers allow us to link companies in Diffbot to those in the Register and to other company-level data, including commercial data products like Orbis Historical and Orbis IP. Second, Diffbot provides extremely rich information on individuals and companies, including detailed education

[5] We query Diffbot for active, for-profit organisations founded before 1 January 2025, and for individuals in employment as of 1 January 2025. Queries run 9 December 2025. In Diffbot Query Language (DQL), our organisation query is *type:Organization not (isDissolved:true) not(isNonProfit:true) NOT(foundingDate>"2025-01-01")*. Our DQL worker query is *type:Person employments.{from <= "2024-12-31"}*.

[6] Our DQL organisation query is: *type:Organization not (isDissolved:true) not(isNonProfit:true) NOT(foundingDate>"2025-01-01") location.country.name:"United Kingdom" has:companiesHouseIds*. Our DQL worker query is: *type:Person location.country.name:"United Kingdom" employments.{from <= "2024-12-31"}*. Queries run 9 December 2025.

and career histories, job titles and descriptions, skills, and companies' most likely key partners, suppliers and competitors. For workers, we use these to enhance the data, building proxies for seniority, migrant status, and to map individual skills. Third, like all web data, Diffbot's core sampling frame is implicit and will not be structured like a conventional sample (Dahlke et al., 2025; Nathan & Rosso, 2022). Unusually, however, Diffbot's data provision is highly transparent, allowing us to see provenance and confidence scores for every element in the graph. This provides crucial insight into sampling frames and helps with validation. We discuss the build in detail in Section 3, and validation in Section 5.

Diffbot's graph updates every four to five days. This means that the characteristics of any sample may change slightly, depending on when it was extracted. This is – very broadly – equivalent to conventional data being revised by statistical authorities in subsequent editions, a common occurrence. In what follows, we timestamp our data on the dates of query and subsequent extraction. In the final build, we include only companies where we observe in Diffbot at least 25% of the stated headcount in Orbis; we remove companies where the share of observed workers was 25% or more when we initially queried the data, but is below 25% when worker profiles are extracted. See Section 3.5 for details.

2.2 Companies House / Open Corporates

Diffbot's implicit sampling frame for organisations and individuals is the public internet. A large literature documents the historically uneven coverage of organisations, countries communities and individuals online (González-Bailón et al., 2014; Graham et al., 2014; Hargittai, 2020). Today's social networks are notably more representative (Chetty et al., 2022; Jeffers, 2024). To understand Diffbot's properties we need to link it to conventional, well-understood datasets where standard financial and other variables are available. We use the UK Companies Register (hence 'Companies House') as the basic sampling frame for companies (and hence their workforces). Companies House data is open, and we use a cleaned, validated version provided pro bono by OpenCorporates, a cross-national compiler of company registry data.[7] All limited companies in the UK need to register at Companies House when they set up, and are provided with

[7] https://opencorporates.com.

a CRN identifier. All overseas companies with a UK branch also need to register, as do some UK business partnerships. As of Q1 2025, the register includes 4.8m active companies, including public and private limited companies, holding companies and partnerships. Of these around 5% will be 'dormant' (no accounts filed for at least 12 months) or non-trading (Companies House, 2025).

All registered companies must file annual returns and annual financial statements ('accounts'), with penalties for non-compliance. Returns cover details of directors and other officers, registered office address, shares and shareholders, company type and (self-assessed) principal business activity. Account detail varies by company size, as defined by turnover, balance sheet and employment thresholds. Companies above these thresholds must file complete, externally audited accounts including revenue and employment, as well as detailed financial information. Given our objective is creating employer-employee data, we restrict our sampling frame to these companies, for whom employment information will be most extensive and high quality. We use the 2016 thresholds to create a time-consistent sample of larger companies with the most complete and highest-quality information. These companies represent a minority of all firms - like most OECD countries, the vast majority of UK firms are small and medium-sized enterprises (SMEs) - but employ the vast majority of the UK's workforce. See Section 3.2 for more details.

2.3 Orbis Historical

OpenCorporates and Companies House do not provide the full contents of company accounts as structured data (only the documents or links to them). We therefore use Bureau van Dijk's Orbis Historical (hence 'OH') to give us detailed financial information on companies over a long timeframe.[8] Orbis includes harmonised cross-country financial information on close to 462 million companies worldwide and is one of the most reliable sources of companies' financial data. Standard Orbis offers financial data over a ten-year timeframe for unlisted companies; Orbis Historical considerably longer. For the UK, Orbis Historical records date back to the mid-1990s - but with considerable attrition pre-2000. We use OH to generate our starting sample of larger

[8] https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html.

companies, using Companies House disclosure thresholds, which we then match to companies in Diffbot. We also use OH to generate information on productivity, company structure and balance sheet characteristics. In both cases, we follow the cleaning procedures for OH data documented in Kalemli-Ozcan et al (2015) and De Loecker, Obermeier and Van Reenen (2024) See Section 3.2 for more details.

2.4 PATSTAT / Orbis IP

In order to supplement company-level information on financial performance with data on innovation, we use PATSTAT Global and Orbis IP to identify patents by companies in our sample. We initially match patents to firms by using Orbis IP, which links companies in OH to their patenting activity using fuzzy matching on company/patent applicant name, address and other observables. We use PATSTAT Global to reconstruct patent families and to build patent quality measures based on citations. See Section 3 for more details.

2.5 Secondary data

We use a number of secondary UK datasets for validation, drawn from the UK Office of National Statistics (ONS). We use the Business Population Estimates (BPE) to benchmark companies in Diffbot (Department for Business and Trade, 2025). We use the 2021 Census and 2022 Labour Force Survey to benchmark Diffbot workforce characteristics (Office for National Statistics, 2024; Office for National Statistics Census Division & Northern Ireland Statistics and Research Agency, 2025). See Section 3 for more details.

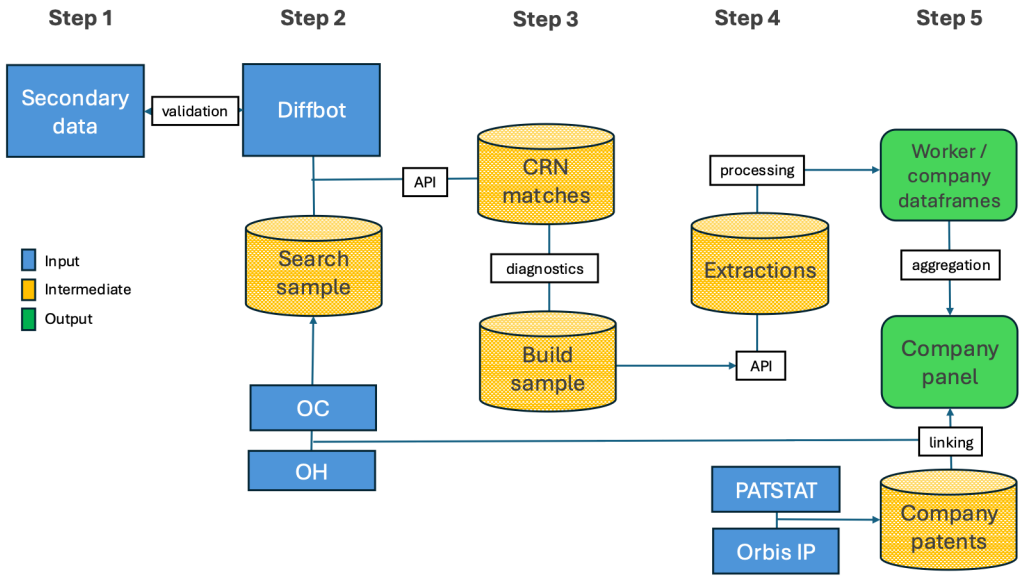
3. Workflow

Figure 2 summarises our workflow. The workflow has five steps. **Step 1** validates Diffbot graph data against secondary sources. **Step 2** involves building the initial company search sample. In **Step 3** we search for these companies in Diffbot, retrieving exact matches and creating a build sample of companies using workforce coverage rate thresholds. In **Step 4** we query Diffbot APIs

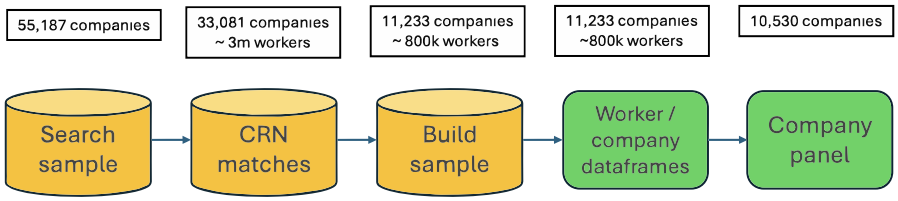
for organisation and worker-level profiles for these companies, clean the extracted data and construct worker-firm dataframes. In **Step 5** we aggregate these data to a company-year panel, and merge in information on firm performance, financials and patents from our other sources.

Figure 2. Workflow overview.

Panel A. Overall workflow



Panel B. Sample size through the build



Source: Authors' elaboration, OpenCorporates, Diffbot, Orbis Historical, PATSTAT, Orbis IP.

3.1 Diffbot sourcing and balance checks

As discussed in Section 2, Diffbot's graph is not a conventional sample of firms or workers, rather the result of continual scraping and learning processes on the public internet. We first explore Diffbot's dominant online sources for UK companies and workers. To understand the

implicit sampling frame, we then compare the distribution of companies and workers in Diffbot against administrative sources.

Key sources: we run two exercises. First, we sample 100 companies from our platform of larger UK businesses - see below for details of the build. We extract the list of online sources and identify domains from full URLs. Second, we repeat the exercise for a sample of 100 UK-based employees of these 100 companies. As expected, Diffbot draws from across the public web, including public social media profiles, business intelligence websites, business directories and the UK's open company repository. For our sample, public profiles from professional networks are the dominant source.

Comparison with administrative data: we compare Diffbot's coverage of UK companies and workers against UK administrative data for which the sampling frame is known. We use Diffbot figures for the start of 2022, to allow us to make worker comparisons with the 2021 England and Wales Census, as well as to labour force and business data. We first compare Diffbot company counts against those in the UK company register, Companies House (CH), and firm counts from the ONS Business Population Estimates (BPE). The intuition for this test is as follows. In CH, each company observation represents a legal entity, not necessarily an actual business: real-world firms may include multiple corporate entities.[9] By contrast, the BPE includes the total number of active private sector businesses, built from the population of actual firms, captured from business tax data, plus an estimate of sole proprietorships. Diffbot takes CH as an input and uses supervised learning to identify the underlying business (an 'organisation' in Diffbot's ontology). Table 2 gives results. Our Encouragingly, we find that our most precise Diffbot specification – counts of active companies in with CRN identifiers – is significantly lower than the count of entities in Companies House, and only slightly larger than the count of enterprises in the BPE.

Next, we compare the characteristics of UK companies and workers in Diffbot across a range of dimensions. Results are summarised in Figure 3. Panel A compares the distribution of active Diffbot companies across industries at the start of 2022, against those in the 2022 BPE.

[9] For example, holding companies exist only to co-ordinate corporate groups, and some firms create separate companies for each function of the business (e.g. HQ, manufacturing, regional distribution and regional sales).

Panel B repeats the exercise for regions. Overall, the distribution of companies across industry and region space approaches those in the BPE. Reflecting the online source data, Diffbot appears to oversample on UK sectors (like financial services) where the majority of firms have an online presence, and undersample on sectors (like construction) where firms typically operate with minimal or zero online presence.

Table 2. Company counts in Diffbot vs UK administrative data, 2022.

Sampling frame	**Diffbot**	**BPE**	**Companies House**
All active for-profit and non-profit organisations	10,588,606		
All active for-profit orgs	10,223,570		
All active for-profit orgs, start of 2022	8,462,847	5,508,935	
All active for-profit orgs, start of 2022, with CRNs	**3,497,151**	**2,947,932**	**5,012,950**

Source: Diffbot, ONS, OpenCorporates. Diffbot results report variations on the query *type:Organization not (isDissolved:true) location.country.name:"United Kingdom"*. Row 1 reports this query. Row 2 adds the condition *not(isNonProfit:true)*. Row 3 adds the condition *NOT(foundingDate>"2025-01-01")*. Row 4 adds the condition *has:companiesHouseIds*. UK Business Population Estimates (BPE) include the total number of private sector businesses, including: companies, all partnerships, and estimates of sole proprietorships (including those unregistered for PAYE/VAT). Estimates in row 3 include all private sector firms at the start of 2022, including companies, all partnerships and sole proprietorships. Estimates in row 4 include companies, sole proprietors with staff and sole proprietors registered for PAYE and/or VAT. Companies House includes all registered companies, including companies, LLPs and some other partnerships. We exclude dormant and non-trading companies.

Panel C compares UK worker age and gender distribution in Diffbot at the start of 2022, against 2021 England & Wales Census data. Diffbot's UK workforce coverage skews young and slightly male. Panel D compares highest reported qualifications for workers in Diffbot against the Labour Force Survey, which provides detailed information on qualifications for a sample of UK-based workers. Diffbot skews significantly upwards on qualifications, with shares of graduates

and postgraduates substantially higher than corresponding LFS estimates. This is consistent with other studies using similar data (Dorn et al., 2025; Fedyk & Hodson, 2022).

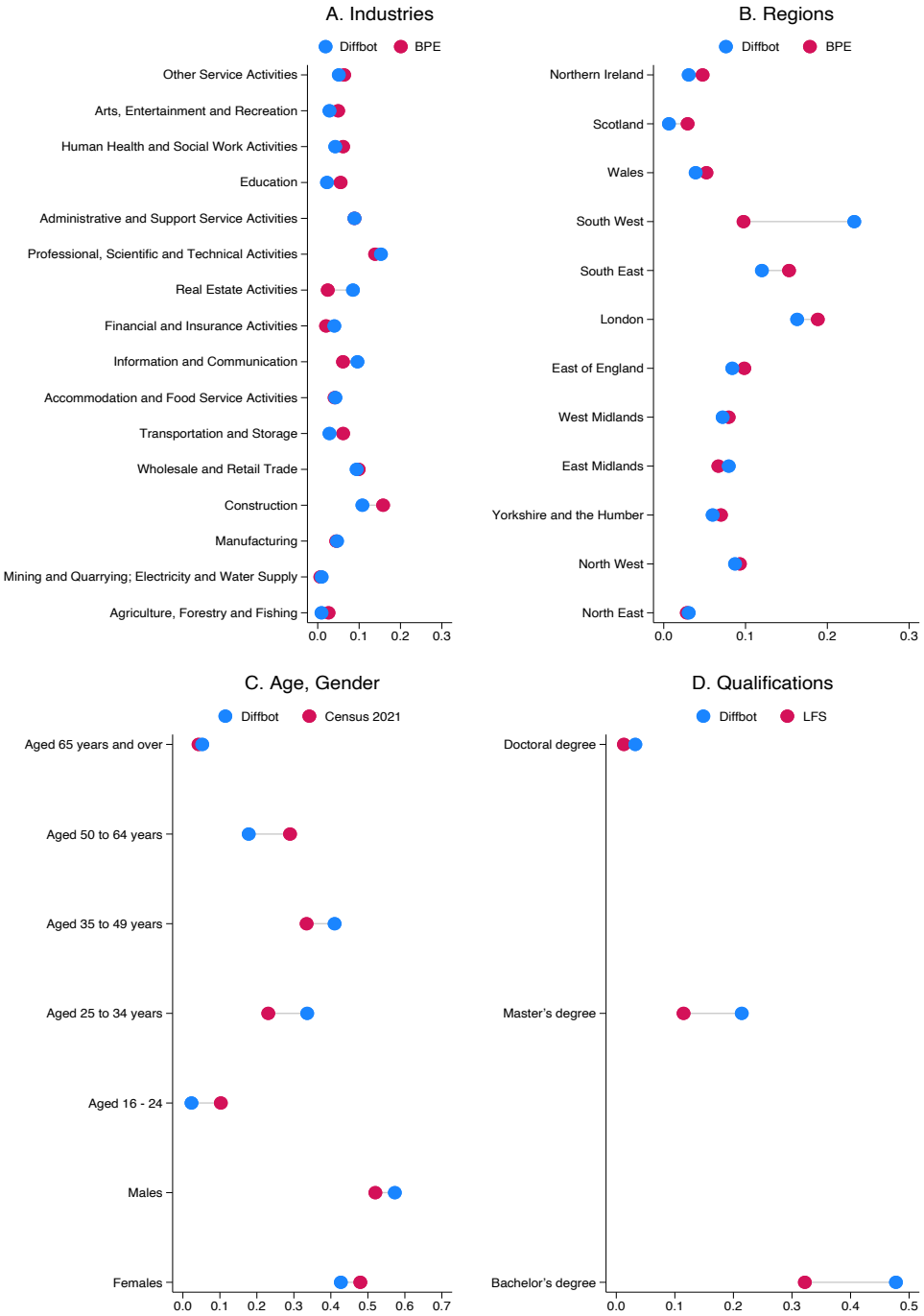
3.2 Company search sample

We construct a sample of 55,187 larger UK companies in the UK active at some point between 2007 and 2023. We use this 'search sample' as the sampling frame for companies in Diffbot. Our company data comes from Open Corporates and Orbis Historical, which provide cleaned, validated versions of data from the UK's company register, Companies House. As explained in Section 2.1, financial and employment information is more complete and higher quality for larger firms. Specifically Companies House distinguishes 'micro' and 'small' companies entities from medium and large companies based on three thresholds based on turnover, assets and employee counts.[10] From 2016, specifically, medium and large companies meet at least two of the following thresholds in any reporting period: 1) Annual turnover more than £10.2 million; 2) Balance sheet total more than £5.1 million; 3) Average annual number of employees more than 50. We apply these post-2016 thresholds backwards to ensure a time-consistent sampling frame.

Choice of timeframe: we start our sample in 2007 for two reasons. First, a key binding constraint is that Orbis Historical data has information on companies' ownership structure only from 2007 onwards. Second, older information from the Web is less accurate for inactive and dissolved firms, e.g., websites are no longer working for dissolved companies, so it is more challenging to assign workers to firms. Companies House, a key source in our case, stops reporting dissolved companies after ten years from the dissolution year. Note that individual worker and firm profiles will often refer to events taking place before this date (for example, educational and career histories, firm formation or lifecycle event).

[10] Thresholds for "medium" sized businesses provide additional exemptions in terms of what is included in the directors' report. However, medium companies are required to file their profit and loss account, and notes to the accounts, so revenues and employment should be available for such firms.

Figure 3. Diffbot benchmarking: organisations and individuals.



Source: Diffbot, ONS, Census, LFS. Panel A shows NACE1 shares for UK-based organisations active in Diffbot (end-2021) and firms in ONS Business Population Estimates (start of 2022). Panel B shows NUTS 1 shares for UK-based organisations active in Diffbot (end-2021) and firms in ONS Business Population Estimates (red, start of 2022). Panel C shows age group and gender shares for individuals in Diffbot employed at the end of 2021 and individuals in the 2021 Census (England and Wales). Panel D shows undergraduate, Masters and PhD shares for individuals in Diffbot employed at the end of 2019, and 2019 LFS data for the UK. Note that for Panel D, Diffbot data is raw, no cleaning.

Approach: we apply larger company thresholds to Orbis Historical data for the period 2007-2023, using data from unconsolidated balance sheets only. We keep only firm-year observations for which financial variables are expressed in pounds sterling. We use the account closing date to determine the calendar year. If the closing date is after or on 1 June, we assign it to the current year. If before 1 June 1, we assign it to the previous year. At this stage, since Orbis may contain multiple annual observations for some firms, we keep the closest observation to the end of the calendar year, which is the annual report in most cases.

These steps give us 55,775 OC-OH companies. We then match this sample of companies to OpenCorporates, which separates company registries by country, to cleanly identify companies incorporated in the UK. This leaves us with a search sample of 55,187 companies observed for 2007-2023.

3.3: Diffbot search, diagnostics and build sample selection

We query the search sample in Diffbot using company identifiers. We return 33,081 companies (~60% of the search sample) with CRN matches. This is our 'match sample' of CRN-linked companies. From the match sample, we select the companies with the highest share of workers present in Diffbot. This 'build sample' consists of 11,233 companies with a coverage ratio of 25% or higher.

Initial search: we search for OC-OH companies in Diffbot. We use Companies House identifiers (CRNs) as our search term and look for the company in each of the years 2007 through 2023 inclusive. Searches were run between April and May 2024. We retrieve the set of companies where the CRN exists in Diffbot and the company has at least one worker observed in Diffbot in any of the sample years. Non-returns indicate that either there is no CRN match, or the company has zero workers observed in Diffbot in the sample years. Specifically, we run a Diffbot Query Language (DQL) call for each company c in our list of CRNs, and each year t (2007-2023). The

query returns everyone working in *c* by the end of year *t*.[11] Note that this initial search cannot directly specify that workers have to be in the UK when working for the firm.[12] Instead, we exploit the fact that the firm has a CRN to narrow down the possible set of workers to those most likely based in the UK at the time of their employment: that is, by conditioning on the firm having a CRN we should almost always return the set of UK-based employment spells. In some cases the query will also pick up non-UK-based workers employed by the UK-based firm; for example, if the firm has a non-UK subsidiary. In the full data, we clean for this by looking at the location of each employment spell. See Step 4 for more detail.

CRN match rate: we find 33,081 companies (59.9% of the OC-OH sample) with at least one employee in Diffbot in any of the search years. The total number of all-time employees observed in Diffbot is 3.04m. On average, we observe just over 55 employees per firm in the sample period (with a standard deviation of 122). Conditioning on matched firms, the average all-time workforce is 91.80 (s.d. = 146.9). Our CRN match rate is well over half, but less than complete. We run diagnostics and a linear probability model on the match sample to better understand match rate predictors. Full results are given in Section 5.

Coverage rate: since Diffbot itself is not a conventional worker-firm sample, the share of the workforce observed in Diffbot will vary - both between companies in the match sample, and within the same company over time. We explore these coverage patterns and use workforce coverage to help us select the final sample of companies. For company *c*, year *t* we define the coverage rate C as:

$$C_{ct} = (\textit{Workers in Diffbot})_{ct} \,/\, (\textit{workers in OC-Orbis})_{ct} \qquad (1)$$

Employment levels often vary substantially year-on-year, as firms lay off workers or scale up (Haltiwanger et al., 2013). Normally researchers adjust for this using a moving average measure;

[11] Necessarily the query includes both people who are still working in *i* and people who stopped working there after year *t*. The query drops anyone who started working in *i* after the end of year *t*; and anyone who stopped working in *i* before the start of *t*. The query also does not return individuals with no employment start/end date.

[12] This is a hard constraint in DQL at the time of writing. The Person endpoint cannot currently accept *employments.{employer.location}* or *employments.{location}* arguments.

in our case, the pairwise correlation between *C* and a two-year moving average is 0.96, so we use the unadjusted measure. We also build the all-time coverage ratio *AC*, which is pooled across our entire sample period.

Coverage rate diagnostics: we run a series of diagnostics on match sample coverage, both on the overall distribution of the coverage ratio, and its breakdown across company age bins, workforce size, industries, regions and performance. In Section 5, we run further validation checks to explore how coverage ratios are affected by companies' observable characteristics and cross-company heterogeneity.

Table 3 shows summary statistics for *C* and *AC*. Panel A shows results for the whole sample; Panel B excludes companies where the coverage rate exceeds 1 (that is, we observe more workers in Diffbot that are reported in Orbis Historical). We discuss reasons for these 'excess' coverage rate cases below.

Table 3. Coverage rate summary statistics.

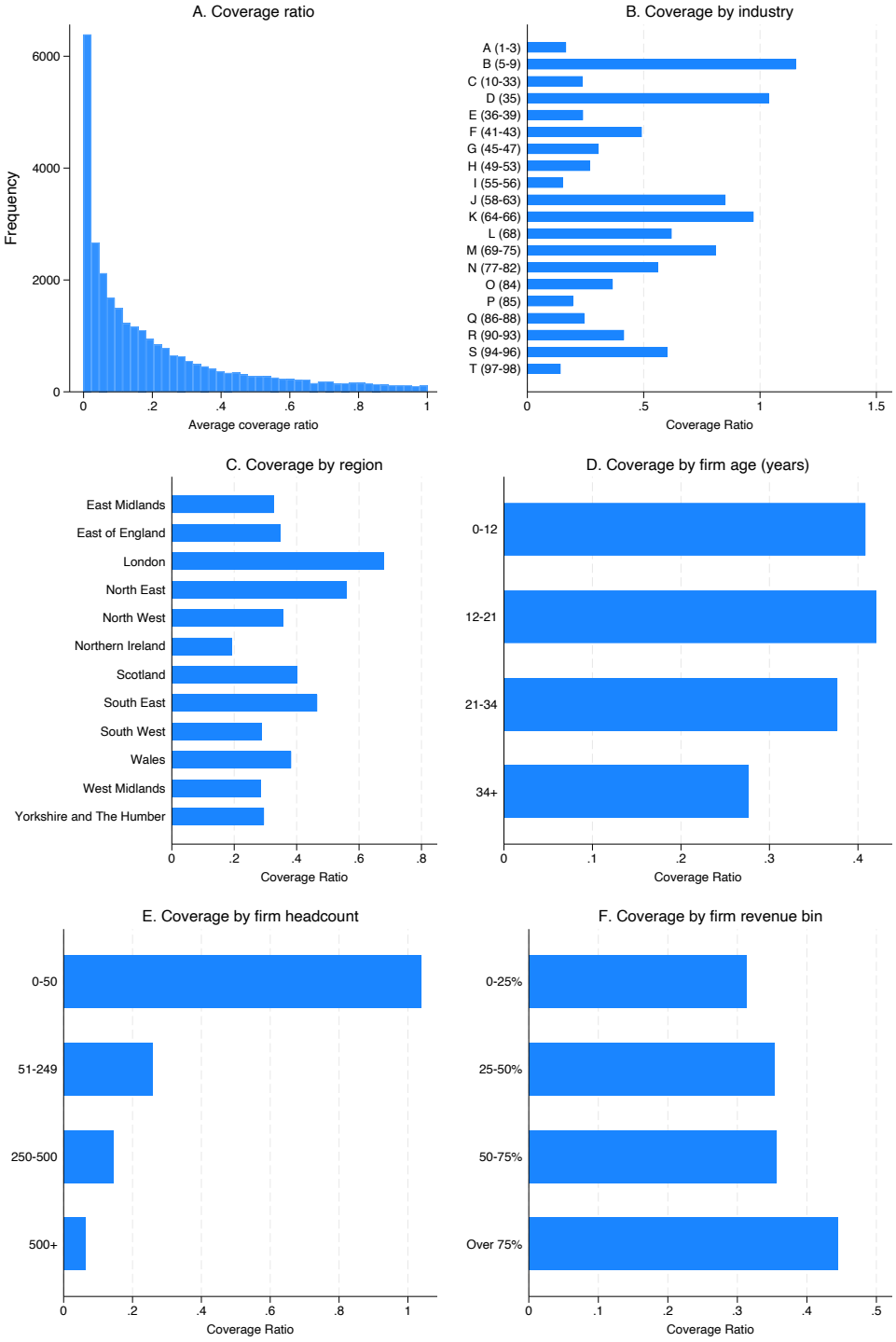
	Median	Mean	25th p	75th p	Obs
Panel A: #Diffbot employees/#Orbis employees					
Coverage rate (C)	0.12	0.37	0.03	0.33	230,981
All-time coverage rate (AC)	0.14	0.44	0.03	0.38	31,347
Panel B: #Diffbot employees/#Orbis employees excluding obs. with coverage ratio > 1					
Coverage rate (C)	0.11	0.19	0.03	0.27	217,413
Average coverage rate (AC)	0.12	0.21	0.03	0.30	28,955

Source: Diffbot, Orbis Historical, OpenCorporates. Notes: The table shows descriptive statistics for the time-varying coverage ratio and the average coverage ratio. N = 33,081 companies.

In the full sample, the median all-time coverage rate is 14%, with a mean of 44%; the interquartile range spans 3-38%. Excluding 'excess' AC cases, the corresponding values are 12%, 21% and 3-30%. We can - roughly - benchmark our sample performance by comparing to Fedyk and Hodson, who provide figures for large US companies matched to Cognism data.(Fedyk & Hodson, 2022) For 2016, their data covers a larger proportion of workers, with a median of 26%, mean of 32% and IQR of 15-42%. Two factors drive these differences. First, both Cognism and Diffbot have better coverage for US businesses and workers than for the rest of the world (in 2016, 34% (36%) of Cognism (Diffbot) employees were in the US/Canada, versus 24% (23.4%) in Europe). Second, per Section 1, Cognism (and its competitors) focus on labour market data and worker profiles, while Diffbot covers a much broader ontology. Diffbot's key advantage for our case is that it provides a direct, precise linkage from workers to firms, while other datasets lack this linkage. In effect, we trade off a lower rate of true positives for a lower rate of false positives. Figure 4 summarises overall workforce coverage (Panel A), as well as coverage across a range of dimensions: 1-digit SIC industry, region, as well as company age, workforce size and reported revenue bins (Panels B - F).

Of the 33,081 companies in the match sample, 21,434 (65%) have an average coverage ratio between 0 and 0,.25; 9,140 (28%) have an AC between 0.25 and 1 and 2,507 (7%) have an AC more than 1 (that is, more workers are observed in Diffbot than are reported in OC-OH). Overall, this is most likely driven by Diffbot ascription error, specifically some combination of: 1) employees are incorrectly assigned to companies with similar names; 2) employees at non-UK branches of a multinational are assigned to the UK parent; 3) non-standard employment (contractors, board members etc.); 4) duplicate worker profiles. None of these can be backed out when using these aggregate data, but we are able to explore and fix all of them when building the individual layer of the data. Specifically, we fix 1) and 2) by looking at worker location, 3) by looking at job title and 4) by de-duplicating individual profiles. For this reason, we keep firms with coverage ratios above 1 in our data at this stage.

Figure 4. Workforce coverage in the match sample.



Source: Diffbot, Orbis Historical, OpenCorporates. Notes: N = 33,081 companies. Average coverage ratios reported, including companies with AC > 1. Industries in Panel B are SIC1 bins. Regions are UK Government Office Regions. Age from incorporation year. Employment is workforce observed in Diffbot. Revenues are quartiles of reported operating revenues.

Build sample selection: our build sample is the 11,233 companies with an average coverage ratio of at least 0.25. These companies have all-time employment information for about 2m workers, of whom we observe around 800,000 in our sample period. We arrive at this threshold as the best trade-off between a high minimum level of coverage and flexibility in subsequent analysis. First, our diagnostics show a clear *J*-curve in coverage ratio, with the bulk of companies having coverage ratios below this (and typically under 10%). Second, per above, benchmarking suggests comparable commercial products have similar coverage ratios when matched to company data. Third, picking a relatively low threshold allows us to flexibly vary coverage quality in regressions, conditioning inclusion on much higher coverage ratios, or weighting observations on coverage ratio without losing observations.

3.4: Diffbot extraction and processing

For the build sample of companies, we extract individual worker and company profiles from the Diffbot Search API. Extracts were run between May and July 2024.

Worker extraction: we extract profiles for the all-time workforce for each company in the final sample. These give us individual characteristics (age, gender, skills etc.) as well as education and employment histories used to build the worker-firm panel. For each company identifier *c* we run a query on Diffbot's Person endpoint, which returns all the person profiles of the all-time workforce. We define this as any worker *i* who 1) is living / has lived in the UK and 2) is working for / has ever worked for company *c*.[13]

Organisation extraction: we also extract company profiles for each company in the final sample. These give us additional textual information on company descriptions and activities, key products and services, key partners / suppliers and competitors, and external finance raised. For each company *c* we return the company profile associated with that company's CRN.

[13] Again, we would ideally specify the employer and employment location, but these arguments cannot currently be passed through Person endpoint queries.

Processing: the API queries return around 800,000 worker JSON objects, plus around 11,200 company JSONs. The JSONS are deeply nested, especially for individual profiles, and we 'flatten' them into CSV format for cleaning. Since Diffbot's key contribution is worker-level information, our focus is on processing workforce data. For companies, we clean each static Diffbot profile and match address information to UK administrative geographies. For workers, we transform each static Diffbot profile into separate dataframes that cover key characteristics, education history, and employment history. Basic characteristics include age, gender, nationality, and languages spoken. Education history variables include start and end date of each education instance, course name, qualification, degree field, institution and institution location (typically city/country pairs). Institutions are coded with Diffbot identifiers which allow us to link to companies/organisations elsewhere in Diffbot. Employment history variables include start and end date of each employment spell, course name, job title, employment location, employer and employer location (usually full addresses or city/country pairs). Employers are coded with Diffbot identifiers, which allow us to link workers to companies/organisations elsewhere in Diffbot. Employment histories include both the companies in our final sample, and any other company/organisation the individual has worked at.

Below we highlight and discuss some key challenges in building the linked worker-firm platform and our design choices. Section 5 sets out validation checks against these choices.

Worker human capital: As discussed in Section 1, worker skills are a central concern for research using worker-firm data. We construct human capital variables analysing text from information on individual education histories. First, we perform an extensive data cleaning process of institution description fields and names, e.g., removing punctuations, standardizing entity names, etc. Second, we created a dictionary of keywords to identify education entities in textual descriptions of institutions, e.g., "school", "college", "university", "diploma", etc. Third, we used this same set of keywords to differentiate between high-school and university degrees and identify an individual's educational attainment.

Worker birth country: Many researchers have used longitudinal employer-employee datasets to study the economic effects of immigration (Foged & Peri, 2016; Kemeny & Cooke,

2018; Malchow-Møller et al., 2013). Diffbot contains fields for person birth country and nationality, but in our data these are almost always blank, reflecting the underlying source data. However, the large majority of people go to secondary school or university in their country of birth. A line with recent papers, we proxy workers' country of birth using the country of their lowest recorded education in Diffbot (Amazadeh et al., 2024; Gupta, 2023; Jin et al., 2025; Lee & Glennon, 2023). This is typically an undergraduate degree, but in about 15% of cases it is high school qualifications (at age 16 or 18). We extract education location information from descriptive free text (e.g. 'University in London, England'); where this is not available we map institution names to UK government dictionaries of UK schools, and a global list of towns and cities with at least 1,000 inhabitants.[14] In theory our measure is vulnerable to false positives (for example, UK-born UK-based workers educated abroad) and false negatives (non-UK born UK-based workers educated here, especially at university level). However, in countries that are net exporters of higher education, like the UK or US, our measure will understate the true number of migrant workers in our sample: in 2022/3, for example, 14.7% of undergraduate students in the UK were non-UK born (Cuibus et al., 2025; HESA, 2024).

Worker language: Alongside skills, mobility and migration, language is another key topic for researchers using employer-employee data (Dale-Olsen & Finseraas, 2022; Ozgen, 2021). We use the first language workers say they speak. Over 70% of workers do not state a given language; as the sampling frame is UK companies, we assume the first language is English. When people give this information in languages other than English, we use a language detection algorithm and assume that the language in which they write is their first language. Adopting the classification made by Ethnologue,[15] we map given first languages to language branches and language families, removing cases where people declare a non-native level of competence.

Worker main jobs and multiple jobs: about 58% of workers in our data have more than one job in any given month. Where the person holds a non-executive / advisory role in a company (e.g. board member, chair), we allow for simultaneous roles since these are typically held in

[14] https://get-information-schools.service.gov.uk/; https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/table/?disjunctive.cou_name_en&sort=name.
[15] https://www.ethnologue.com.

combination. These cases comprise 11% of simultaneous job holders. For the other 89%, we interpret overlapping positions as an unobservable combination of genuine workers in two or more part-time positions; plus noise from errors in online resumes and/or old employers not updating staff pages post-move. In these cases we force a single role by assigning the worker to the firm they have worked for in a past spell (59% of the remainder); if they have no employment history in any firm in multiple job scenario, we assign them to the most senior job on the basis of job title text (25.7%); if this is not possible we use random assignment (4.3%).

Worker occupations: understanding changes in occupational structure is a central topic for workers using employer-employee data: for example, in studying firms' adaptation to technological change, or the evolution of labour markets over time (Acemoglu, Anderson, et al., 2023; Acemoglu, Koster, & Ozgen, 2023). Diffbot employment histories include workers' job titles. We use LinkTransformer to crosswalk these to UK 2020 Standard Occupational Codes (SOC2020). LinkTransformer is an open-source LLM-based classifier optimised for matching text (Arora & Dell, 2023). Specifically, we match the universe of Diffbot job descriptions in our worker data to SOC2020 descriptors. We experiment with different levels of SOC granularity, achieving the best performance with 4-digit descriptors. Validation checks confirm the quality of the match; see Section 5.

Worker skills: Researchers have long used linked employer–employee data to examine how workers' qualifications and years of experience shape earnings, and influence firm. However, the skills that arise from study or learning by doing are rarely observed directly and at scale.(Dorn et al., 2025) GRAPH-EE enhances this line of inquiry by incorporating granular skill information drawn from Diffbot's knowledge graph. Diffbot includes a typology of 32,000 skills, validated as meaningful using prevalence in Wikidata and online professional platforms. Input text is drawn from person profiles (for example, job titles, job descriptions, self-ascribed skills and endorsements from others). Intuitively, we can think of Diffbot skills as representing learned capabilities that help workers complete tasks. As explained by Dorn and co-authors, in human capital models workers acquire skills through formal education or on the job experience (Becker, 1962; Dorn et al., 2025). In practice, workers with the same qualifications may acquire quite distinct skills based on tasks done through their careers (Autor, 2013; Autor & Dorn, 2013; Autor

& Handel, 2013). A worker's level of Diffbot skills should then be positively correlated with experience and be distinct from formal qualifications. In Section 5, we directly test these predictions for our data.

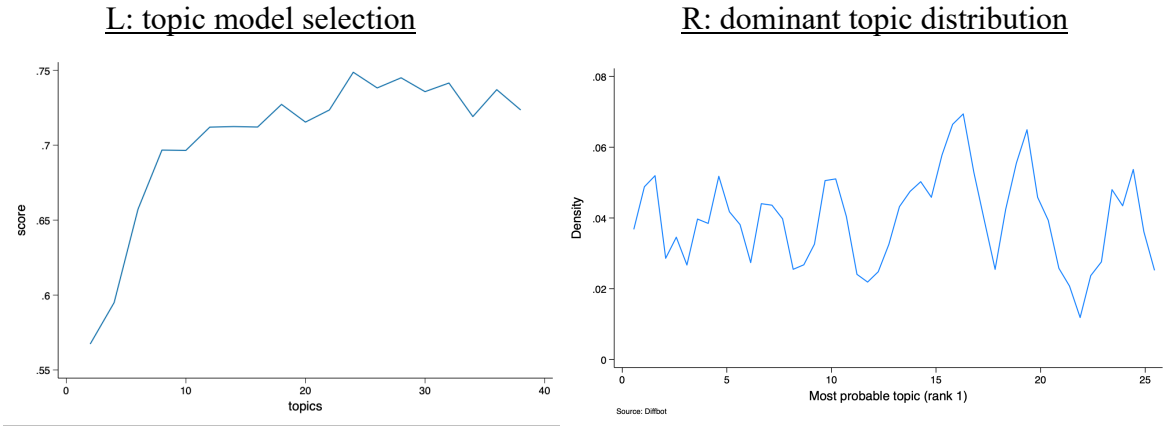
We briefly discuss other potential concerns here. First, Diffbot's skill standardisation workflow provides some reassurance that we are not picking up meaningless or idiosyncratic content. Second, workers without skills may differ from those with skills, for example in their willingness to report skills, or the willingness of others to endorse them. Over 80% of the workers in our data - approximately 655,000 individuals - have observed skills. In Section 5 we run tests for selection into skills, finding no evidence of selection on observables, and evidence that unobservables play very little role. Third, note that skills are only observed for the last year in which a person appears in Diffbot. One-third are observed in 2023; the rest (5-10% in each year) are observed between 2014 and 2022. We treat our data as either cross-sectional or pooled cross-sections based on the distribution of worker tenure in our data, as explained below.

To make the raw skills data tractable, we use topic modelling to reduce skills dimensionality. We select a 25-topic model based on goodness of fit (see Figure 5). Topics are summarized using the top 10 and top 300 words. We use the Large Language Model GPT-4o to label each topic. We also ask the LLM to use ISCO descriptors to rank topics by the complexity of the implied tasks, hence 'ISCO-complexity'. We use GPT-3.5 and Claude to provide alternate labellings and rankings (see Section 5 for results). The topic modelling algorithm assigns topic probabilities to individuals based on their observed skills, with probabilities summing to one. For each individual, we retain the three topics with the highest probabilities; we refer to the topic with the highest probability as the 'dominant topic' for that individual. Figure 5 shows the fit metric for model selection (left panel) and dominant topic prevalence (right panel). In the preferred 25-topic model dominant topics are evenly distributed across workers.

We make company-level metrics for the final panel. These cover workforce average skill complexity; share of workers with 'high-complexity skills' (the top 8 / 25 topics); and a Fractionalisation Index of skills across all 25 topics. These allow researchers to explore technical specialisation and skill diversity respectively. We treat skills topics as cross-sectional,

representing the skills workers have in the job in which they are observed. In our data, mean job tenure is 4.5 years and median tenure is three years. We therefore build firm-level skills metrics for 2023, the last year of our data, and as averages in two to five-year windows between 2019-2023 inclusive.

Figure 5. Diffbot skills topic modelling.



Source: Diffbot. Sample is 655,504 workers with skills observed between 2007 and 2023.

Worker and company location: employer-employee data is widely used to explore the economic growth of cities and regions, leveraging information on firm and worker location (Card et al., 2025; Combes et al., 2008; De La Roca & Puga, 2017; Koster & Ozgen, 2021). GRAPH-EE currently locates focal companies at city level, and worker employment histories at country level. For workers, we retrieve country location by fuzzy-matching employer names in Diffbot back to the universe of companies in Orbis Historical. Future versions of the data should be able to provide full histories at city level – in the raw data historical employer city location is missing in 45.4% of cases, but could be backfilled using Diffbot organisation profiles from further API extractions. For companies, based on validation exercises (see Section 5), we match companies in our sample to the Travel to Work area (TTWA) and UK regions given by their 'current main address' in Diffbot. We prefer this to registered addresses in Companies House, which are typically given at incorporation and may not reflect the location of trading activity – we show in Section 5 that this is still a non-trivial concern for the larger companies in our sample. TTWAs represent functional labour markets akin to city-regions, or US Commuting Zones. Regions are larger administrative geographies akin to US States. Note that while we cannot precisely locate multi-site companies

from company-level information, we can infer site locations at city/country level from worker location information.

Step 5: Company panel

Step 4 produces a series of worker- and company-level data frames. To build the company panel, we create a single worker * company * year matrix, aggregate this to the company * year level, then merge in company information on financials from Orbis Historical, as well as innovation measures from PATSTAT / Orbis IP.

Productivity measures: following the literature, we build two core productivity measures. One is a standard measure of Total Factor Productivity, which represents the contribution to the firm's output that is not accounted for by capital or labour. First, to maximise coverage, we construct Value Added as the sum of earnings before interest, taxes, depreciation and amortization (EBITDA) and costs of employees (Bajgar et al., 2020; De Loecker et al., 2024). Labour input is measured by the number of employees, capital input by fixed assets. Second, we derive Total Factor Productivity (TFP) by estimating the following log-transformed production function, across industries, for company *i*, industry *j,* year *t*:

$$Y_{ijt} = \alpha + \beta 1 L_{ijt} + \beta 2 K_{ijt} + e_{ijt} \quad (2)$$

where Y is log(output), L is log(labour) and K is log(capital). Then, TFP is estimated as a residual as follows:

$$TFP_{ijt} = \hat{\alpha} - Y_{ijt} - (\widehat{\beta 1} \cdot L - \hat{\beta} 2 \cdot K)_{ijt} \quad (3)$$

where $\widehat{\beta 1}$ and $\hat{\beta} 2$ are the industry-specific elasticities obtained from estimating (2) for each industry bin. We implement using standard non-parametric production function estimators (Ackerberg et al., 2015; Olley & Pakes, 1996).

Company innovation measures: we use patents to measure companies' innovative activities. Patents are the most established proxy for innovation, specifically inventive activity, and have well-known advantages and drawbacks (Castaldi et al., 2020; Hall et al., 2014; Nathan & Rosso, 2022). We first assign patents to firms using ORBIS IP, then we reconstruct patent families using PATSTAT Global. A patent family is a set of patents applied to different patent offices covering the same underlying invention. Reconstructing patent families is essential to avoid that the same invention is counted more than once. We construct two measures of innovation: quantity, the weighted count of patents for each company (where jointly held patents are weighted by the number of applicants); and quality, the lifetime count of forward citations on each patent. We assign patents to the priority year of application, considered the closest date to the original invention.

Final company panel: this panel combines aggregated company-year information from Diffbot with the productivity and innovation measures above. We first aggregate Diffbot worker-firm-year data to company-year level, giving us a panel of 11,233 companies. Linking this to companies' financial data gives us an unbalanced panel of 10,530 companies observed between 2007 and 2023. Note that this includes some observations with missing TFP, location or industry fields, as well as around 10% of companies where the workforce converge ratio at extraction stage is less than 25%. We keep these in our data to allow researchers maximum flexibility.

4. Data Records

A company-level public-release version of GRAPH-EE is available through the UK Data Service ReShare platform, collection ID 858769. This data comprises the company-level aggregate outputs from the build, minus variables taken from Orbis Historical / Orbis IP. This reflects licensing constraints imposed by Bureau van Dijk. The data does not include individual-level variables developed from Diffbot data. This reflects our data-sharing agreement with Diffbot. Our codebase will allow anyone with access to Orbis and Orbis Historical, Orbis IP/PATSTAT Global/equivalent, and Diffbot APIs to reproduce our complete worker-firm data. Table 4 summarises the main features of the public-release data and the complete data.

Table 4. Overview of the final data.

Dataset	Scale/s	Key fields	Sources
Public	Company	CRN Birth year Main 4-digit industry Current main full address + TTWA + region Registered address + TTWA + region Workforce coverage ratio Share graduates Share post-graduates Share PHDs Share Oxbridge graduates Share Russell Group grads Share arts and humanities / social sciences / economics / STEM graduates Share migrant; graduate migrant Share female; graduate female Share managerial / tech / STEM roles Share workforce with high/medium/low skills Workforce average skills	Diffbot OpenCorporates
Complete	Company	Above, plus BvD ID Total Factor Productivity (TFP) Gross Value Added (GVA) Operating revenue EBITDA Number of workers Other financials / capital / investment and costs Cost of employees Number of subsidiaries Foreign subsidiaries dummy Other Diffbot organisation fields All-time patent counts, citations	Diffbot OpenCorporates PATSTAT Global Orbis Historical Orbis IP
Complete	Worker	Education history (start and end date, institution, subject, qualification) Employment history (start and end dates, employer, location, job title) Worker skill topics, ranked and unranked	Diffbot

5. Technical Validation

On top of our validation of Diffbot sourcing and benchmarking against administrative sources (Step 1 in our workflow), we validate the quality of the dataset through two further types of exercise. First, we check what drives CRN matching and variations in workforce coverage, core elements in constructing the company build sample (Step 3). Second, we validate the construction of a number of key variables in the subsequent worker-firm panel build (Step 4).

5.1 Match and coverage rates

To validate the construction of the search and build samples, we run checks on predictors of a) the match rate between Diffbot and the search sample, and b) the workforce coverage ratio for the set of CRN matches that are used in the build sample.

<u>CRN matching checks</u>

We precisely match 60% of our search sample of companies with organisations in Diffbot. Diffbot scrapes Companies House profiles and identifiers, so in theory every company in our search sample should precisely match to a Diffbot organisation. We therefore need to explore determinants of the match rate. We first manually explore Diffbot sources for a sample of 100 non-matches. This check shows Diffbot's workflow scrapes Companies House, but then sometimes fails to place CRN data in the relevant organisation profile cell. This pattern of ascription error appears to be random.

Next, we formally test whether company observable characteristics might plausibly influence Diffbot's matching workflow. Specifically, for all 55,187 companies in the search sample (Step 2), we estimate a linear probability model where the dependent variable is a dummy equal to one if the company is CRN-matched in Diffbot and zero otherwise. The set of predictors include a dummy for whether the company is dissolved, the log of the average company size (measured by the number of employees from Orbis), incorporation year, log of the average total assets, log of the average revenues, average number of subsidiaries, a binary variable for whether

the company has non-UK subsidiaries, a dummy for whether the company is a Global Ultimate Owner (GUO) of its group, plus a set of 1-digit SIC industry dummies and a set of macro-region dummies. More formally, we estimate the following linear probability model for company *i*:

$$Pr(match)_i = F(\boldsymbol{Observables})_i \quad (4)$$

Where ***Observables*** is the set of predictors described above, plus 1-digit industry dummies and region dummies. As matches are non-time-varying we estimate (6) cross-sectionally.

Table 5 gives results. Overall, our results suggest that CRN (non)-matching is not explained by company characteristics: predictor effect sizes are small, model fit is very low, and unobservables play little role in explaining remaining variance. The most important statistically significant predictor, being a dissolved company, reduces the probability of a match by just over 2%. We also show that unobserved company characteristics do not drive this result, using an Oster Test on the dissolved dummy.[16] The test parameter indicates the required magnitude of any unobservables, relative to observables, to cancel out the effect of interest. In our case, unobservables need to be 43 times larger than observables to nullify the effect we find. This seems highly implausible.

[16] As recommended by Oster (2019), we set a maximum R^2 equal to 1.3 times that obtained in the specification with all control variables and fixed effects.

Table 5. Match rate predictors test.

Dependent variable: CRN match in Diffbot	**(1)**
Dissolved	-0.212***
	(0.00667)
Average company size	0.0171***
	(0.00193)
Incorporation year	-0.00107***
	(0.000110)
Log assets	-0.0392***
	(0.00194)
EBITDA	0.00136***
	(0.000225)
Number of subsidaries	-0.000439
	(0.000526)
Foreign subs.	-0.100***
	(0.0357)
Company is Global Ultimate Owner	0.0928***
	(0.00533)
Industry FE	Yes
Region FE	Yes
Observations	47576
R^2	0.0736
Oster delta, dissolved	43.01

Source: Diffbot, Orbis Historical, OpenCorporates. Notes: The table shows the results of a linear probability model of the probability of a company matching on CRN in Diffbot, on company observables characteristics, industry and region dummies. The dependent variable is a dummy that equals one if the company is matched in Diffbot on CRN, zero otherwise. Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Coverage rate checks

We do not observe all workers in a given company. Our all-time workforce coverage rate varies between 21% and 44%, and this figure will also vary across and within companies over time. To explore coverage rate predictors, we adapt a test by Fedyk and Hodson, who explore coverage of Cognism for a sample of large US firms (Fedyk & Hodson, 2022). Specifically, we estimate

$$Coverage_{ijt} = F(\boldsymbol{Observables}_{ijat}, I_i, J_j, A_a, T_t) \quad (5)$$

Where *Coverage* is the coverage ratio of company *i*, sector *j*, region *a* in year *t,* ***Observables*** are the same as (4), and I, J, A and T are company, sector, region and year fixed effects. We first run pooled OLS regressions for *C* and *AC*. Results are given in Table 6.

Column (1) looks at predictors of the annual coverage ratio; column (2) the average coverage ratio across all years of our data. In both cases, while some selection on observables is present, our results suggest it is trivial: some coefficients are significant, but effect sizes remain small and model fit is very low. To test for the influence of unobservables on workforce coverage we run an Oster Test on the most important predictor, in this case firm size. The delta is between 3.1 and 4.3, suggesting that unobservables would need to be three to four times more important to dominate our main result. Column (3) extends the test to a typical employer-employee setting, where researchers fit company and year fixed effects. When we do this, we can explain almost 80% of the coverage ratio variation in terms of observables and fixed effects. In the published data, we include coverage ratios by company and year. This allows researchers to flexibly control for coverage in a variety of ways, for example including *C* as a control variable; weighting observations by coverage, or subsetting to higher-coverage companies.

Table 6. Coverage ratio predictors test.

Dependent variable = workforce coverage ratio	(1)	(2)	(3)
Dissolved	-0.0329***	-0.0817***	
	(0.0169)	(0.0158)	
Firm size	-0.350***	-0.288***	-0.752***
	(0.0194)	(0.0157)	(0.0475)
CH Incorporation year	0.000835***	0.00128***	
	(0.000163)	(0.000147)	
Log assets	0.153***	0.131***	0.175***
	(0.0150)	(0.0131)	(0.0174)
EBITDA	-0.000417	-0.000761	-0.000317
	(0.000505)	(0.000476)	(0.000308)
Number of subsidiaries	-0.000672	-0.00178*	0.00410***
	(0.000981)	(0.000942)	(0.000912)
Foreign subsidiaries	-0.183***	-0.0919**	0.0338
	(0.0445)	(0.0449)	(0.0222)
Company is Global Ultimate Owner	0.0104*	0.00346	-0.00490
	(0.00630)	(0.00553)	(0.0119)
Industry FE	Yes	Yes	Yes
Region FE	Yes	Yes	Yes
Firm FE	No	No	Yes
Year FE	Yes	Yes	Yes
Observations	184096	184096	180752
R^2	0.0444	0.0437	0.791
Oster delta, firm size	2.904	4.164	

Source: Diffbot, Orbis Historical, OpenCorporates. Notes: The table shows the results of OLS regressions of companies' workforce coverage ratios on company observable characteristics, industry and region dummies and firm fixed effects. Dependent variables are the company coverage ratio *C* (columns (1 and 3) and the company average coverage ratio *AC* (column 2). Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5.2 Worker-firm build

We run a series of validation checks on key worker and company level variables.

Worker education and birth country

As explained in Section 3, for around 10% of workers, education location is missing and this needs to be imputed using a dictionary method. Correct imputation is important, not least because we impute worker birth country based on the country of lowest observed education. As discussed above, this is a design used in a number of other papers – but which also needs validation, and testing the prediction that it is likely to provide a lower bound on the true migrant share.

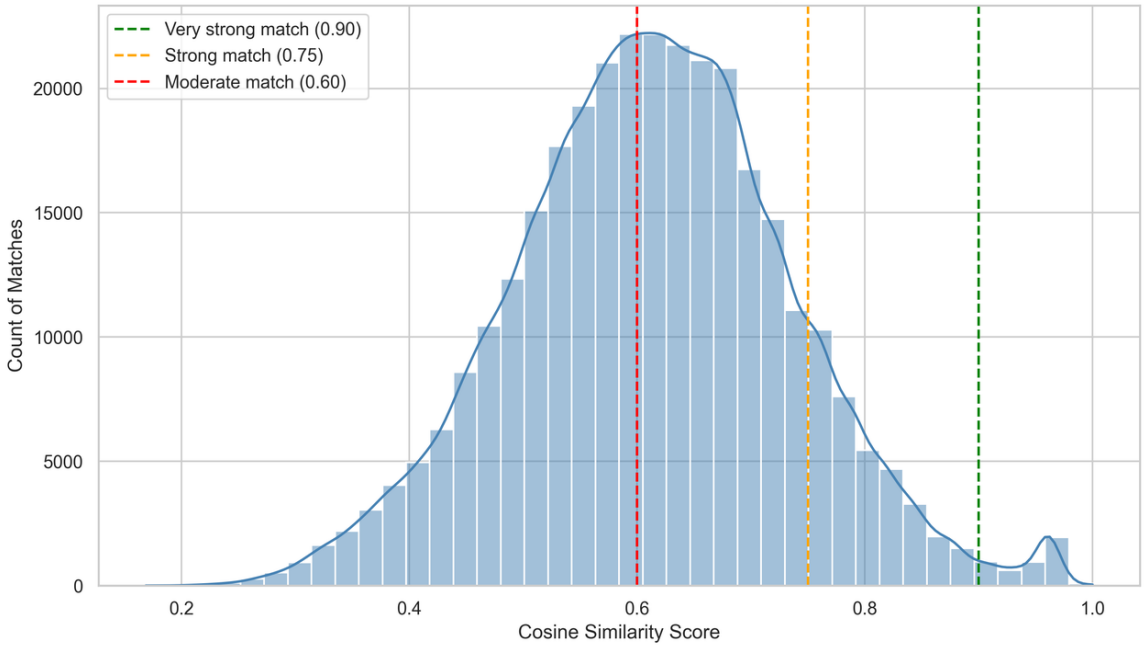
We ran two checks on our processed data. First, to test the quality of imputation an RA manually validated a random sample of 100 worker observations where we have directly observed levels and locations of education spells. We impute education location/s and compare to the observed information. Country of education is correctly imputed for 90/100 observations. Second, we test the country of lowest education proxy on a sample of staff and PhD students from a UCL Department, a setting a) where we have ground truth and b) individuals are disproportionately highly qualified, mirroring the larger sample of workers in Diffbot. We use a simple web survey, obtaining a response rate of 35%. All of our respondents attended university. Birth country correctly maps to country of university for 89.7% of respondents. Results for country of schooling are identical. As predicted, country of university predicts a lower bound for the true migrant share (51.4% versus 54.3%).

Worker occupations

We use LinkTransformer to match job descriptions in Diffbot to UK Standard Occupational Code descriptors. Both SOC descriptors and Diffbot job titles are first standardised through a uniform text-cleaning pipeline that removes non-informative tokens, employer references, and formatting artefacts. Matching is then performed within LinkTransformer using its default sentence-embedding model (all-MiniLM-L6-v2), producing one-to-one matches between

cleaned job titles and SOC descriptors. The approach delivers high precision at fine occupational granularity, performing particularly well at the 4-digit (SOC Unit Group) level. In-sample matching scores exhibit a smooth, unimodal distribution with substantial mass at high similarity values, indicating that most assignments correspond to semantically close matches rather than spurious links. See Figure 6.

Figure 6. Distribution of LinkTransformer matching scores.



Further, an RA validated the match for 100 job descriptions sampled from workers in our final panel. 85% of the match job descriptions were scored as correct at the 4-digit level. Of the rest, 14% were scored incorrect, and 1% unable to score based on SOC descriptor text.

Worker skills

In Section 3, we suggested that Diffbot skills are a proxy for learned capabilities that derive from experience and are distinct from formal qualifications. Given that many people work in fields unrelated to formal qualification, this implies that for any given type of job, skills should be a) positively correlated with workers' years in the labour market, and b) weakly correlated with levels of qualifications. We confirm both predictions in Table 7. We build a matrix of 4-digit SOC bins

and show pairwise correlations of SOC-level average job complexity (Panel A) or shares of 'high-skill' jobs (Panel B) against that SOC's share of graduates, share of post-graduates, share of PHDs and mean workforce experience in years. Both panels show a negative link between workers' Diffbot skills and the share of graduates in a given job, but weaker positive correlations with postgraduate and PHD shares, and strong positive correlations with workforce experience. Working with similar data for US graduates, Dorn and co-authors show that most worker profiles report skills that are plausibly related to experience, and that the number of content of skills are positively correlated with years of experience and with wages (Dorn et al., 2025).

Table 7. Benchmarking Diffbot skills against qualifications and experience.

Panel A.	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
GPT-4o topic ranking	1.000				
Share graduate	-0.360***	1.000			
Share post-graduate	0.047	0.164***	1.000		
Share PhD	0.210***	-0.136***	0.419***	1.000	
Ave worker experience	0.350***	-0.068	-0.008	-0.003	1.000
Panel B.	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
(mean) high_skill	1.000				
Share graduate	-0.236***	1.000			
Share post-graduate	0.066	0.164***	1.000		
Share PhD	0.234***	-0.136***	0.419***	1.000	
Ave worker experience	0.243***	-0.068	-0.008	-0.003	1.000

Source: Diffbot. Notes: Table shows correlation matrices for a sample of 412 SOC4 bins for 655,000 workers with Diffbot skills. Panel A shows the pairwise correlation of average worker skill complexity in each SOC against shares of graduates / post-graduates / PhD workers in the SOC, and average worker experience in the SOC. Panel B repeats for SOC shares of workers with 'high-skill' dominant topics (topics 17-25 inclusive). * p<0.1, ** p<0.05, *** p<0.01.

Next, we run checks against skills coverage and skills topic ranking. Just under 20% of workers in our sample have no skills information in Diffbot. We run a linear probability model regressing a worker's probability of having Diffbot skills, controlling for individual characteristics,

the kind of job they are doing at the time, and the year skills are observed in the data. For worker *i* in 4-digit occupation bin *o* observed in year *t,* we regress

$$\Pr(Y_{iot} = 1) = F(\boldsymbol{X}_{iot}, O_o, T_t) \quad (6)$$

Where Y is a dummy taking the value one if a worker has observed skills, **X** is a vector of worker observables, O is one of 411 SOC4 fixed effects and T is the year skills are observed, from 2007 to 2023.

Results are shown in Table 8. Overall, model fit and coefficient effect sizes are low. Speaking a foreign language and having a PhD, the most important predictors, increase the probability of having skills by around 7% and 3% respectively. Oster tests on these variables give deltas of 8.8 and 4.1 respectively, suggesting unobserved worker characteristics are essentially trivial in explaining whether or not Diffbot skills are observed.

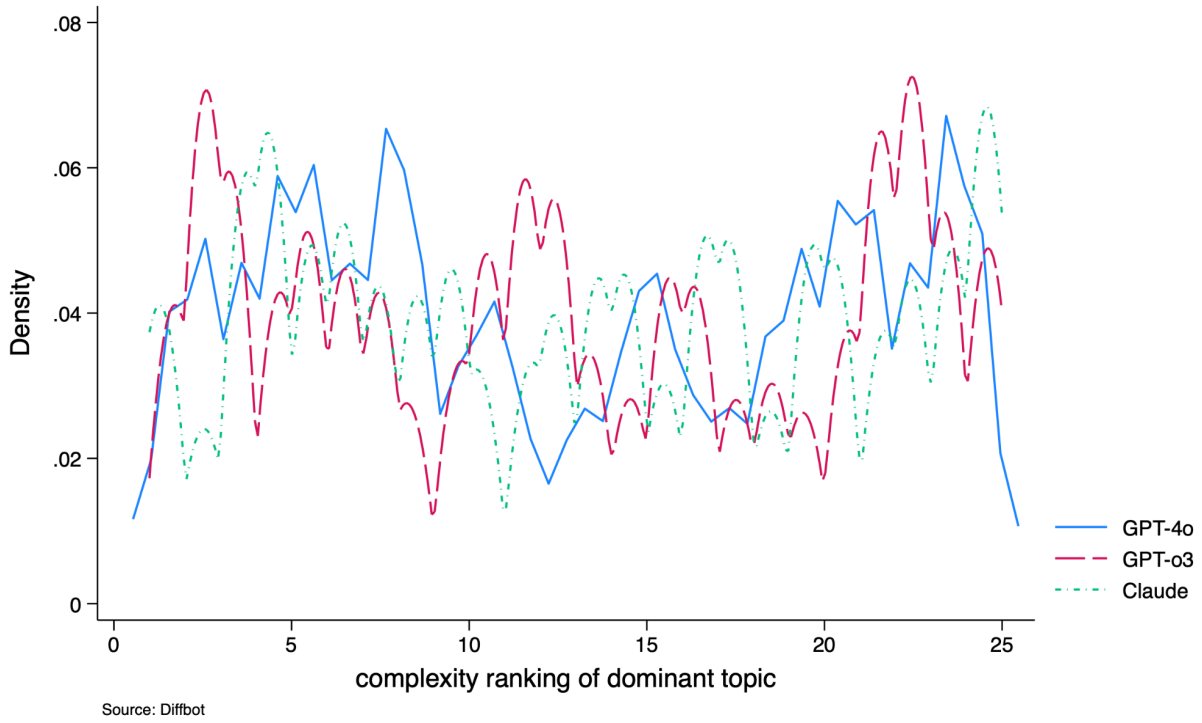
Table 8. Diffbot skills selection test.

Depvar = has Diffbot skills	(1)
Migrant	0.0191***
	(0.00117)
Speaks foreign language	0.0722***
	(0.00125)
Age	-0.00196***
	(0.0000331)
Female	0.00853***
	(0.00101)
Has STEM degree	0.0177***
	(0.00113)
Has economics degree	0.0181***
	(0.00112)
Has arts / humanities degree	0.0252***
	(0.00126)
Studied at Oxbridge	-0.0153***
	(0.00259)
Studied at Russell group uni	-0.000124
	(0.00127)
Graduate	-0.00298**
	(0.00140)
Has postgrad degree	0.00841***
	(0.00175)
Has PhD	0.0308***
	(0.00319)
Yech occupation	0.0219***
	(0.00171)
Managerial occupation	0.0242***
	(0.00121)
Years of experience	0.00935***
	(0.0000708)
Observations	540652
R^2	0.0782
Oster delta, foreign language / PhD	8.882 / 4.114

Source: Diffbot, Orbis Historical, OpenCorporates. Notes: The table shows the results of an LPM regression of worker probability to have Diffbot skills on observable characteristics. All regressions include dummies for year skills are observed, and SOC4 dummies. Standard errors in parentheses clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01.

Finally, we run sensitivity tests on topic ranking, by comparing rankings across GPT-4o, our main classifier, with rankings from two alternate LLMs, OpenAI's GPT-3.5 and Anthropic's Claude. The overall pairwise correlation between LLM rankings is very high: for GPT-4o and GPT-o3 it is 0.865, and for GPT-4o and Claude it is 0.844. Figure 6 breaks this down across topics, showing the distribution of rankings for each LLM.

Figure 6. Distribution of topic rankings by main classifier and alternatives.



Source: Diffbot. Sample is 655,504 workers with skills observed between 2007 and 2023.

Company location

We use Diffbot's 'current primary address' variable to locate companies. To check the accuracy of this information, a research assistant manually compared current primary addresses for a sample of 100 companies with the 'true' trading address, derived from either firms' websites or when unavailable, Google Maps. The RA scored location accuracy at different scales: full postcode, postcode district, city, region and country. Diffbot addresses matched true addresses at

the postcode / postcode district level for around 50/100 of cases; at the city and region level the match is around 80/100.[17] This implies that in the vast majority of cases, locating companies at TTWA or region level will place their main site correctly in physical space. We also find that for the same sample, Diffbot current primary addresses and Companies House registered addresses are identical in at least 75% of cases, even at the full postcode level.

5. Discussion

The full version of GRAPH-EE can be used to explore a range of questions about UK firm performance (including on key metrics such as productivity, revenue and innovation); business dynamics, including the growth of emerging industries; workforce characteristics; and labour market progression and outcomes, including the role of worker skills, qualifications and characteristics. It can also be used for spatial analysis, including exploring industry location patterns and clusters, including in emerging sectors; and workforce skills and characteristics across UK towns and cities.

Caveats: this version of GRAPH-EE consists of company-level layers from OpenCorporates, Diffbot and PATSTAT. Data provider restrictions mean we cannot release variables derived from Orbis Historical, or individual-level records from Diffbot. The code provided should enable users with access to Diffbot's API, and to Orbis Historical, to reproduce our data. Per section 2, note that Diffbot continually updates its graph, and therefore data constructed later than GRAPH-EE may differ slightly from previous versions. Note that our full data includes some observations with missing TFP, location or industry fields, as well as around 10% of companies where the workforce converge ratio at extraction stage is less than 25%. We keep these in our data to allow researchers maximum flexibility.

[17] 28/100 Diffbot addresses have no postcode information; a handful of addresses have no city/region/country information.

References

Acemoglu, D., Anderson, G., Beede, D., Buffington, C., Childress, E., Dinlersoz, E., Foster, L., Goldschlag, N., Haltiwanger, J., Kroff, Z., Restrepo, P., & Zolas, N. (2023). Advanced Technology Adoption: Selection or Causal Effects? *AEA Papers and Proceedings*, *113*, 210-214. https://doi.org/10.1257/pandp.20231037

Acemoglu, D., & Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. In C. David & A. Orley (Eds.), *Handbook of Labor Economics* (Vol. Volume 4, Part B, pp. 1043-1171). Elsevier. https://doi.org/http://dx.doi.org/10.1016/S0169-7218(11)02410-5

Acemoglu, D., Koster, H. R. A., & Ozgen, C. (2023). Robots and Workers: Evidence from the Netherlands. *National Bureau of Economic Research Working Paper Series*, *No. 31009*. https://doi.org/10.3386/w31009

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification Properties of Recent Production Function Estimators. *Econometrica*, *83*(6), 2411-2451. https://doi.org/https://doi.org/10.3982/ECTA13408

Amazadeh, N., Kermani, A., & McQuade, T. (2024). *Return Migration and Human Capital Flows*.

Arora, A., & Dell, M. (2023). *LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models* (ArXiv 2309.00789, Issue.

Autor, D. H. (2013). The "task approach" to labor markets: an overview. *Journal for Labour Market Research*, *46*(3), 185-199. https://doi.org/10.1007/s12651-013-0128-z

Autor, D. H., & Dorn, D. (2013). The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*, *103*(5), 1553-1597. https://doi.org/doi: 10.1257/aer.103.5.1553

Autor, D. H., & Handel, M. J. (2013). Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics*, *31*(S1), S59-S96. https://doi.org/10.1086/669332

Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2022). Artificial Intelligence, Firm Growth and Product Innovation. *Journal of Financial Economics*.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., & Timmis, J. (2020). *Coverage and representativeness of Orbis data* (OECD Science, Technology and Industry Working Papers 2020/06, Issue.

Barreto, C., Damas de Matos, A., & Hijzen, A. (2025). *Immigrant integration: The role of firms* (International Migration Outlook 2025, Issue.

Becker, G. (1962). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Boustan, L., Jensen, M. F., Abramitzky, R., Jácome, E., Manning, A., Pérez, S., Watley, A., Adermon, A., Arellano-Bover, J., Åslund, O., Connolly, M., Deutscher, N., Gielen, A. C., Giesing, Y., Govind, Y., Halla, M., Hangartner, D., Jiang, Y., Karmel, C., . . . Zohar, T. (2025). Intergenerational Mobility of Immigrants in 15 Destination Countries. *National Bureau of Economic Research Working Paper Series*, *No. 33558*. https://doi.org/10.3386/w33558

Breithaupt, P., Hottenrott, H., Rammer, C., & Römer, K. (2025). Linked Employer–Employee Data from XING and the Mannheim Enterprise Panel. *Jahrbücher für Nationalökonomie und Statistik*, *245*(6), 689-703. https://doi.org/doi:10.1515/jbnst-2024-0070

Card, D., Rothstein, J., & Yi, M. (2025). Location, Location, Location. *American Economic Journal: Applied Economics*, *17*(1), 297–336. https://doi.org/10.1257/app.20220427

Castaldi, C., Block, J., & Flikkema, M. J. (2020). Editorial: why and when do firms trademark? Bridging perspectives from industrial organisation, innovation and entrepreneurship. *Industry and Innovation*, *27*(1-2), 1-10. https://doi.org/10.1080/13662716.2019.1685376

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., . . . Wernerfelt, N. (2022). Social

capital I: measurement and associations with economic mobility. *Nature*, *608*(7921), 108-121. https://doi.org/10.1038/s41586-022-04996-4

Combes, P.-P., Duranton, G., & Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, *63*(2), 723-742. https://doi.org/10.1016/j.jue.2007.04.004

Companies House. (2025). *Incorporated companies in the UK January to March 2025*. https://www.gov.uk/government/statistics/incorporated-companies-in-the-uk-january-to-march-2025/incorporated-companies-in-the-uk-january-to-march-2025

Criscuolo, C., Gal, P., Leidecker, T., & Nicoletti, G. (2021). *The human side of productivity* (OECD Productivity Working Papers, No. 29, Issue. https://www.oecd-ilibrary.org/content/paper/5f391ba9-en

Cuibus, M., Walsh, P. W., & Němeček, F. (2025). *Student migration to the UK*. https://migrationobservatory.ox.ac.uk/resources/briefings/student-migration-to-the-uk/

Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., Schütz, M., Kriesch, L., Hottenrott, H., Kanilmaz, U. N., Grashof, N., Hajikhani, A., Liu, L., Riccaboni, M., Balland, P.-A., Wörter, M., & Rammer, C. (2025). *The WebAI Paradigm of Innovation Research: Extracting Insight From Organizational Web Data Through AI* (ZEW Discussion Paper 25-019, Issue.

Dale-Olsen, H., & Finseraas, H. (2022). *Linguistic Diversity and Workplace Productivity* (IZA Discussion Paper 12621, Issue.

De La Roca, J., & Puga, D. (2017). Learning by Working in Big Cities. *The Review of Economic Studies*, *84*(1), 106-142. https://doi.org/10.1093/restud/rdw031

De Loecker, J., Obermeier, T., & Van Reenen, J. (2024). Firms and inequality. *Oxford Open Economics*, *3*(Supplement_1), i962-i982. https://doi.org/10.1093/ooec/odad097

Department for Business and Trade. (2025). *Business population estimates for the UK and regions 2025: statistical release*. https://www.gov.uk/government/statistics/business-population-estimates-2025/business-population-estimates-for-the-uk-and-regions-2025-statistical-release

Department for Education, HM Revenue and Customs, Department for Work and Pensions, & Higher Education Statistics Agency. (2025). *Longitudinal Education Outcomes SRS Iteration 2 Standard Extract - England*. https://doi.org/https://doi.org/10.57906/pzfv-d195

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). *Knowledge vault: a web-scale approach to probabilistic knowledge fusion* (Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14), Issue.

Dorn, D., Schoner, F., Seebacher, M., Simon, L., & Woessmann, L. (2025). *Multidimensional Skills on LinkedIn Profiles: Measuring human capital and the gender wage gap* (Rockwool Foundation Berlin Discussion Paper 136/25, Issue.

Douglas Heaven, W. (2020). *This know-it-all AI learns by reading the entire web nonstop* (Technology Review, Issue. https://www.technologyreview.com/2020/09/04/1008156/knowledge-graph-ai-reads-web-machine-learning-natural-language-processing/

Fedyk, A., & Hodson, J. (2022). Trading on Talent: Human Capital and Firm Performance*. *Review of Finance*. https://doi.org/10.1093/rof/rfac068

Fetzer, T., Lambert, P. J., Feld, B., & Garg, P. (2024). *AI-generated production networks: measurement and applications to global trade.* (Warwick economics research papers series (WERPS) (1528), Issue.

Foged, M., & Peri, G. (2016). Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data. *American Economic Journal: Applied Economics*, *8*(2), 1-34.

Gagliardi, L., Mariani, M., & Breschi, S. (2024). Temporal Availability and Women Career Progression: Evidence from Cross-Time-Zone Acquisitions. *Organization Science*, *35*(6), 2178-2197. https://doi.org/10.1287/orsc.2023.17685

Goldin, C., Kerr, S. P., Olivetti, C., & Barth, E. (2017). The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census. *American Economic Review*, *107*(5), 110–114. https://doi.org/10.1257/aer.p20171065

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16-27. https://doi.org/https://doi.org/10.1016/j.socnet.2014.01.004

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers*, *104*(4), 746-764. https://doi.org/10.1080/00045608.2014.910087

Gupta, A., (May 16, 2023). . (2023). *Labor Mobility, Entrepreneurship, and Firm Monopsony: Evidence from Immigration Wait-Lines* (Available at SSRN: https://ssrn.com/abstract=4450105 or http://dx.doi.org/10.2139/ssrn.4450105 Issue.

Hall, B. H., Helmers, C., Rogers, M., & Sena, V. (2014). The Choice Between Formal and Informal Intellectual Property: A review. *Journal of Economic Literature*, *52*(2), 375-423.

Haltiwanger, J., Jarmin, R., S. , & Miranda, J. (2013). Who Creates Jobs? Small versus Large versus Young. *The Review of Economics and Statistics*, *95*(2), 347-361. http://ideas.repec.org/a/tpr/restat/v95y2013i2p347-361.html (The Review of Economics and Statistics)

Haltiwanger, J., Lane, J. I., & Spletzer, J. R. (1999). Productivity differences across employers: The roles of employer size, age, and human capital. *American Economic Review*, *89*(2), 94-98.

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10-24. https://doi.org/10.1177/0894439318788322

HESA. (2024). *Higher Education Student Statistics: UK, 2022/23 - Where students come from and go to study*. https://www.hesa.ac.uk/news/08-08-2024/sb269-higher-education-student-statistics/location

Heyman, F., Sjöholm, F., & Tingvall, P. G. (2007). Is there really a foreign ownership wage premium? Evidence from matched employer–employee data. *Journal of International Economics*, *73*(2), 355-376.

Jeffers, J. (2024). The Impact of Restricting Labor Mobility on Corporate Investment and Entrepreneurship. *Review of Financial Studies*, *37*(1), 1-44.

Jin, Z., Kermani, A., & McQuade, T. (2025). *Native-Immigrant Entrepreneurial Synergies* (NBER Working Paper 33804, Issue. http://www.nber.org/papers/w33804

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2015). How to Construct Nationally Representative Firm Level Data from the Orbis Global Database: New Facts and Aggregate Implications. *National Bureau of Economic Research Working Paper Series*, *No. 21558*. https://doi.org/10.3386/w21558

Kemeny, T., & Cooke, A. (2018). Spillovers from Immigrant Diversity in Cities. *Journal of Economic Geography*, *18*(1), 213–245.

Kemeny, T., Haus-Reve, S., Dahl-Fitjar, R., & Cooke, A. (2020). Does assimilation shape the economic value of immigrant diversity? *Economic Geography*.

Koster, H. R. A., & Ozgen, C. (2021). *Cities and Tasks* (IZA Discussion Paper 14231, Issue. https://EconPapers.repec.org/RePEc:iza:izadps:dp14231

Lee, S., & Glennon, B. (2023). *The Effect of Immigration Policy on Founding Location Choice: Evidence from Canada's Start-up Visa Program* (NBER Working Paper 31364, Issue. http://www.nber.org/papers/w31634

Malchow-Møller, N., Munch, J. R., Seidelin, C. A., & Skaksen, J. R. (2013). Immigrant workers and farm performance: Evidence from matched employer-employee data. *American Journal of Agricultural Economics*, *95*(4), 819-841.

Mesquita, F., Cannaviccio, M., Schmidek, J., Mirza, P., & Barbosa, D. (2019). *KnowledgeNet: A Benchmark Dataset for Knowledge Base Population* (Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Issue. K. Inui, J. Jiang, V. Ng, & X. Wan. https://aclanthology.org/D19-1069/

Nathan, M., & Rosso, A. (2022). Innovative events: product launches, innovation and firm performance. *Research Policy*, *51*(1), 104373. 10.1016/j.respol.2021.104373

Office for National Statistics. (2024). *Labour Force Survey. [data series]. 11th Release*.

Office for National Statistics. (2025a). *Annual Survey of Hours and Earnings, 1997-2024: Secure Access. [data collection]. 26th Edition, SN: 6689*.

Office for National Statistics. (2025b). *Employment in the UK: April 2025*. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/employmentintheuk/april2025

Office for National Statistics Census Division, & Northern Ireland Statistics and Research Agency. (2025). *2021 Census: Aggregate Data. [data collection]*. .

Olley, G. S., & Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, *64*(6), 1263-1297. https://doi.org/10.2307/2171831

Ozgen, C. (2021). The economics of diversity: Innovation, productivity and the labour market. *Journal of Economic Surveys*, *35*(4), 1168-1216. https://doi.org/https://doi.org/10.1111/joes.12433

Rock, D. (2019). *Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence* (Working paper, Issue.

Tambe, P. (2014). Big Data Investment, Skills, and Firm Value. *Management Science*, *60*(6), 1452-1469. https://doi.org/doi:10.1287/mnsc.2014.1899

Vilhuber, L. (2018). *LEHD Infrastructure S2014 files in the FSRDC* (Center for Economic Studies Discussion Papers 1(2), 3., Issue.