Appendices

A Multiscale community detection with Markov Stability

The skills graph $\mathcal{G}$ is analysed using Markov Stability (MS), a multiscale community detection framework that uses graph diffusion to detect communities in the network at multiple levels of resolution. For a fuller explanation of the ideas underpinning the method, see Refs. [1–4].

Let A be the $N \times N$ adjacency matrix and D be the diagonal degree matrix of a graph $\mathcal{G}$. The transition probability matrix M of a discrete-time random walk on $\mathcal{G}$ is:

$$M = D^{+}A, \tag{1}$$

where D^{+} denotes the pseudo-inverse of D. The matrix M defines a discrete-time Markov chain on the nodes of $\mathcal{G}$ [5]:

$$\mathbf{p}_{r+1} = \mathbf{p}_r\, M \tag{2}$$

where $\mathbf{p}_r$ is a $1 \times N$ vector with components denoting the probability of the random walk arriving at the respective node at discrete time r.

There are different continuous-time processes associated with the random walk [3,6]. In particular, consider the rate matrix

$$Q = M - I, \tag{3}$$

where $I_{N\times N}$ is the identity matrix. Note that $L = -Q$ is the *random walk Laplacian*. We then define the continuous-time Markov process with semi-group $P(r)$ governed by the forward Kolmogorov equation

$$\frac{dP}{dr} = P\,Q, \tag{4}$$

which has the solution

$$P(r) = e^{rQ}. \tag{5}$$

which, under broad assumptions, converges to a unique stationary distribution $\boldsymbol{\pi}$, defined by

$$\boldsymbol{\pi} = \boldsymbol{\pi}\, M, \tag{6}$$

where $\boldsymbol{\pi}$ is a $1 \times N$ probability vector, which fulfills $\boldsymbol{\pi} L = 0$ [7].

Markov Stability as a cost function for clustering algorithms

The dynamics of the Markov chain with transition matrix M defined on the nodes of the graph can be exploited to get insights into the properties of the graph G itself. Following [1, 6], each partition of the graph into c communities corresponds to a $N \times c$ indicator matrix H where $H_{ij} = 1$ if node i is part of community j and $H_{ij} = 0$ otherwise. We can then define the clustered autocovariance matrix for partition H as

$$\mathcal{K}_r(H) = H^T \left[\Pi P(r) - \boldsymbol{\pi}^T\boldsymbol{\pi}\right] H. \tag{7}$$

The diagonal elements $\mathcal{K}_r(H)_{ii}$ correspond to the probabilities that the Markov process starting in one community i does not leave the community up to time r, whereas the off-diagonal elements correspond to the probabilities that the process has left the community in which it started by time r. It is important to remark that r is an intrinsic time of the Markov process that is used to explore the graph structure and is clearly distinct from the physical time of some applications. To avoid confusion, it is customary in Markov Stability analysis to refer to r and/or $s = \log_{10}(r)$ as the Markov *scale*. Following these observations, we define the Markov Stability of a partition H by

$$\mathcal{R}_r(H) = \min_{0\le l\le r} \mathrm{Tr}\left[\mathcal{K}_l(H)\right] \approx \mathrm{Tr}\left[\mathcal{K}_r(H)\right]. \tag{8}$$

The approximation in (8) is supported by numerical simulations that suggest that $\mathrm{Tr}\left[\mathcal{K}_r(H)\right]$ is monotonically decreasing in r. The Markov Stability $\mathcal{R}_r(H)$ is thus a dynamical quality measure of the partition for each Markov scale r which can be maximised to determine optimal partitions for a given graph and each scale of the associated Markov process.

The objective is therefore to find a partition $H(s)$ that maximises Markov Stability up to a time horizon (scale) s for the Markov process on the graph:

$$\mathcal{R}_s\left(H(s)\right) = \max_{H} \mathcal{R}_s(H). \tag{9}$$

Optimisation of Markov Stability for different Markov scales s then leads to a series of partitions $H(s)$. For small Markov scales, the Markov process can only explore local neighbourhoods, which leads to a fine partition, whereas increasing the Markov scale widens the horizon of the Markov process so that larger areas of the graph are explored, which leads to coarser partitions [6]. Hence, the notion of a community as detected by Markov Stability analysis is strictly based on the spread of a diffusion on the network.

B Skill clusters at coarse resolution (MS7)

One of the features of our multiscale analysis is the possibility of extracting clusterings at different levels of resolution that are inherently robust in the data. In our MS analysis (Fig 2 in the main text), we found a robust coarser partition (MS7) into seven skill clusters. Here we cover succinctly some of the findings for this skill clusters following the same procedure and format as for MS21 above.

Table 1: Summary of statistics for coarse resolution skill clusters (MS7).

Index	Cluster	No. of Skills	No. of Mentions	Average Mention	Semantic Similarity	Skill Containment	Closeness Centrality	Average Salary
1	Business and Financial Management	611	217631667	3.348	0.169	0.521	0.142	32600
2	Engineering and Operations Management	1184	117181750	1.803	0.137	0.435	0.144	30680
3	Technical and Software Development	816	113008925	1.739	0.145	0.577	0.130	39060
4	Sales and Marketing	457	88738239	1.365	0.185	0.352	0.137	29310
5	Teaching and Healthcare	688	56098105	0.863	0.139	0.356	0.134	29510
6	Hospitality and Food Industry	85	18441250	0.284	0.179	0.193	0.137	22000
7	Information Security and Cybersecurity	65	2903836	0.045	0.246	0.164	0.119	44110

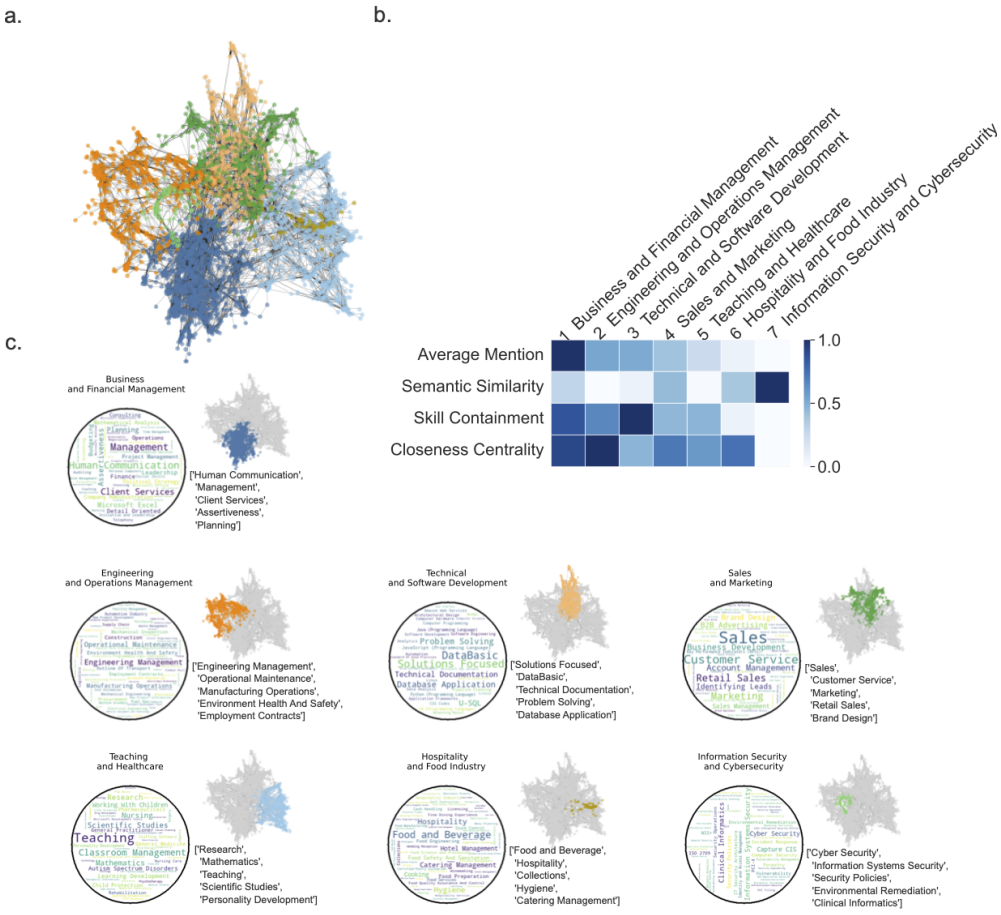

Fig 1: **Co-occurrence skill clusters (MS7)** (a) Skills network coloured according to the 7 skill clusters. (b) Summary heatmap of skill clusters properties. (c) For each of the 7 clusters, word cloud where font size represents skill eigenvector centrality, and list of top 5 most frequent skills.

Table 1 and Fig 1 present a summary of the results with computed statistics for the clusters, summary labels, and word clouds.

Unlike MS21, we see greater imbalance in the number of skills contained in each cluster, with a twenty-fold difference between the cluster with the fewest skills ('Information Security and Cybersecurity', 65 skills) and the most skills ('Engineering and Operations Management', 1184 skills). Also, despite containing half as many skills, 'Business and Financial Management' skills are mentioned

twice as much (3.3 mentions per advert) than 'Engineering and Operations Management' (1.8 mentions per advert). As expected, the within-cluster semantic similarity is broadly lower than in the MS21 configuration, reflecting the larger size of these clusters, which combine skills more different from one another. The 'Information Security and Cybersecurity', the smallest cluster, has higher semantic similarity than the others, with an ostensibly similar skills profile. Indeed, the Sankey diagram in Fig 3 in the main text shows that this cluster is persistent across a range of scales, suggesting these skills occupy a distinct, isolated region in the skills network.

Compared to MS21, skill containment is generally higher, largely as a result of the MS7 clusters containing more skills. The highest containment is found in 'Technical and Software Development', where 57.7% of all co-occurrences between skills involve skills from the same cluster. Conversely, as in MS21, 'Information Security and Cybersecurity' remains poorly contained (16.4%) and is instead often connected to the rest of the skills network, yet with high semantic similarity and low closeness centrality. This reinforces the role of this skill cluster as a specialist cluster yet with broad co-occurrence. The differences in skill centrality across clusters also indicate the relevance of skills in specialised sectors, which are less shared across job adverts (i.e., low centrality) and with high containment, such as 'Technical and Software Development'. See Fig 2 for more details of closeness centrality, skill containment and semantic similarity for the clusters in MS7.

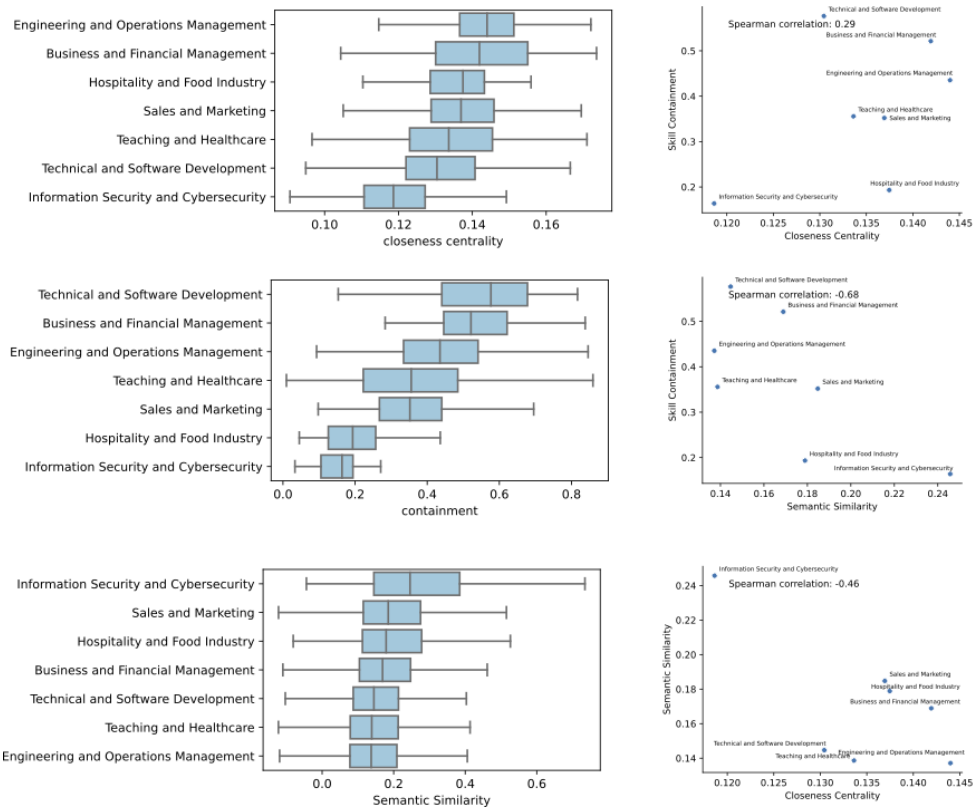

Fig 2: Boxplots for distributions of closeness centrality, skill containment and within cluster semantic similarity for each MS7 skill cluster. The scatter plots compare these three variables.

This extensive overlap of skills into more sectorial groupings also leads to a convergence of the average salary, with 4 clusters having an average salary close to £30,000. The three outlying clusters are 'Hospitality and Food industry' (£22,000) and 'Technical and Software Development' (£39,060) and 'Information Security and Cybersecurity' (£44,100), which can be viewed as specialist clusters with different levels of pay.

Fig 3 presents the geographical distribution of the skill clusters in MS7 across NUTS2 regions. As for MS21, we observe substantial geographical variability that reflects different socio-economic factors, including industrial composition. This will be the object of future work.

Finally, Fig 4 shows the Sankey diagram between the coarser MS7 skill clusters and the 32 LC skill

categories. As for MS21, we find broad agreement, with some differences partly reflecting differences in the scale of the two categories (7 vs 32). For instance, 'Technical and Software Development' in MS7 is closely related to the LC 'Information Technology' category, but also incorporates the LC 'Analysis' category. The LC 'Health Care' category is included almost entirely within the MS7 'Teaching and Healthcare' cluster. Similarly, the MS7 'Sales and Marketing' cluster maps to the LC 'Design', 'Sales', 'Media and Communications' and 'Marketing and Public Relations' clusters. Note also that, unsurprisingly, as the MS clusters become coarser, the overall semantic similarity decreases since more dissimilar skills are grouped together (MS21: 0.172, MS7: 0.169, MS4: 0.141).

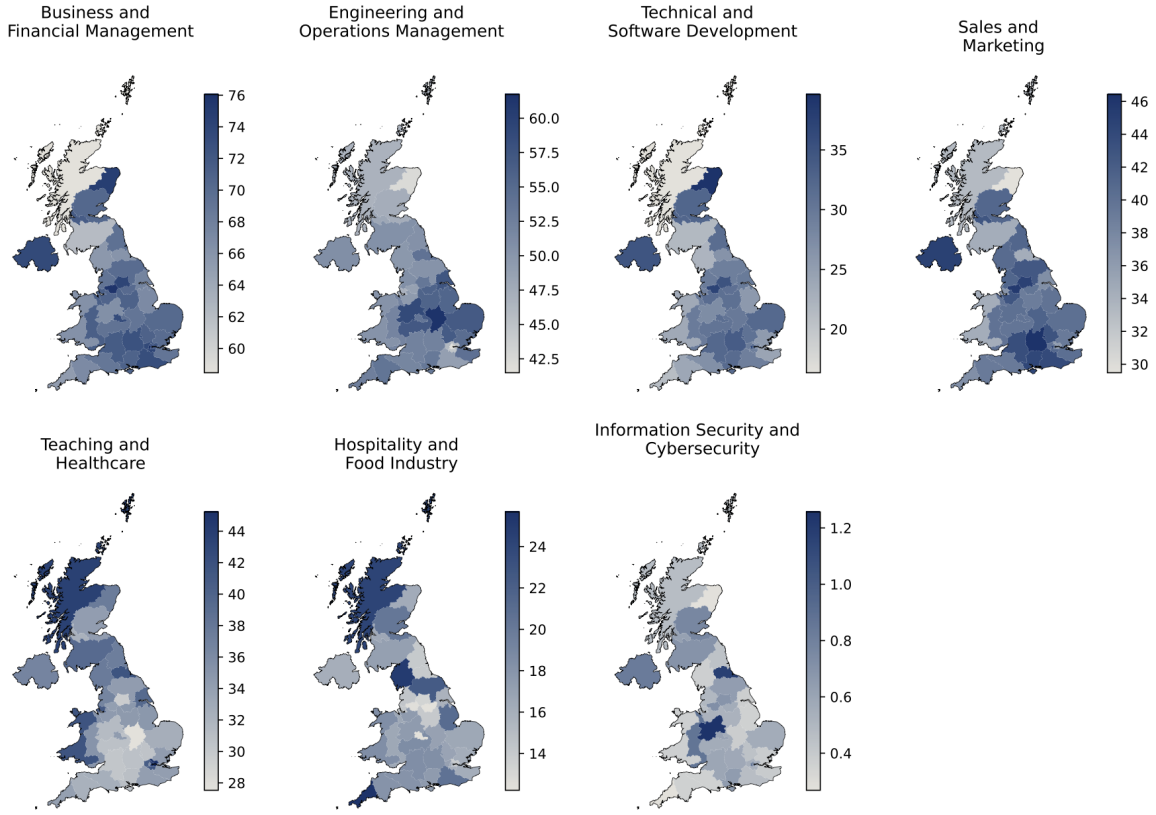

Fig 3: Maps for each of the MS7 clusters showing the percentage of all adverts in each NUTS2 regions featuring a skill from the cluster. Map shapefile source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right [2024].

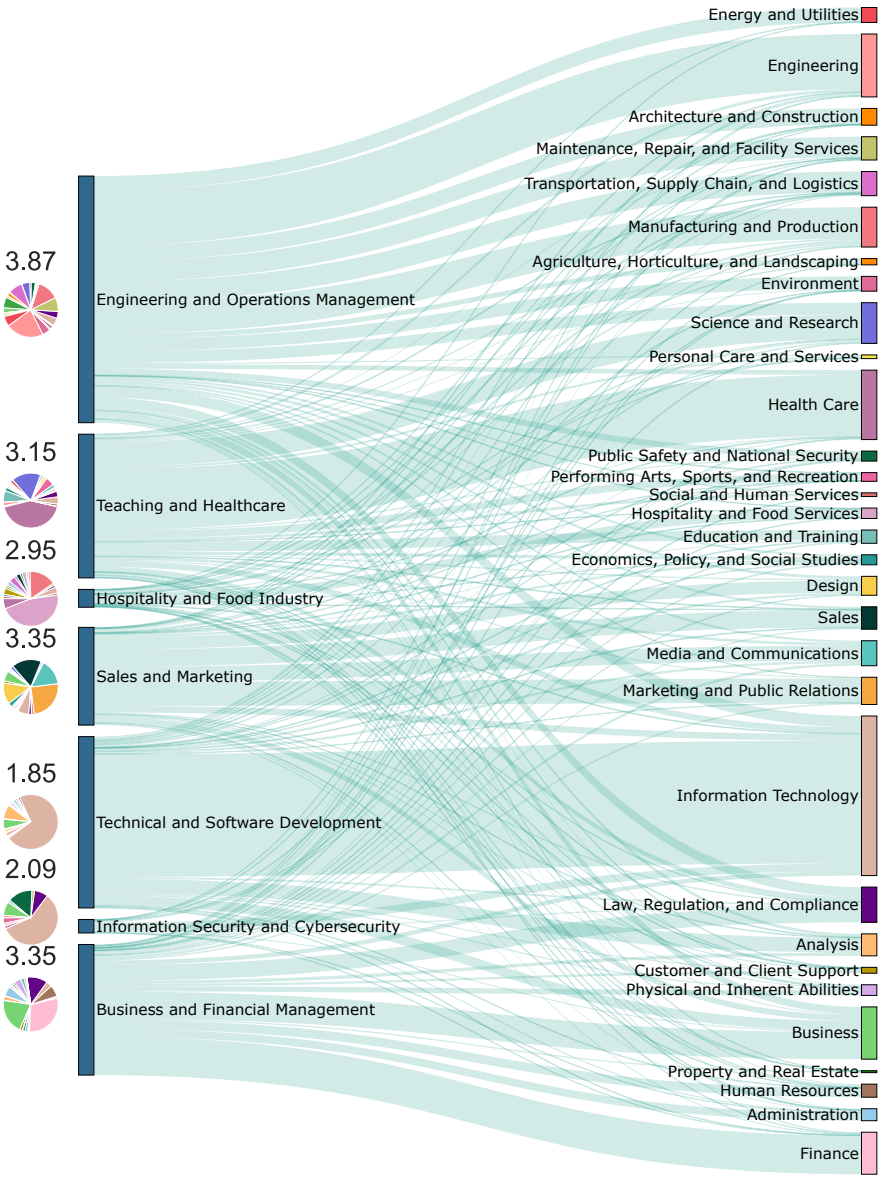

Fig 4: **Sankey diagram between co-ocurrence skill clusters (MS7) and expert-based skill clusters (Lightcast)**. There is broad agreement between the data-driven clusters and the LC categories in some skill areas where thematic content and co-occurrence match. For each cluster in MS7, we plot a pie chart to visualise the proportions of Lightcast categories, and the corresponding entropy to indicate how thematically mixed the cluster is.

C Supplementary analyses

C.1 Percentage of occupations in Adzuna data set relative to official Labour Force Survey statistics

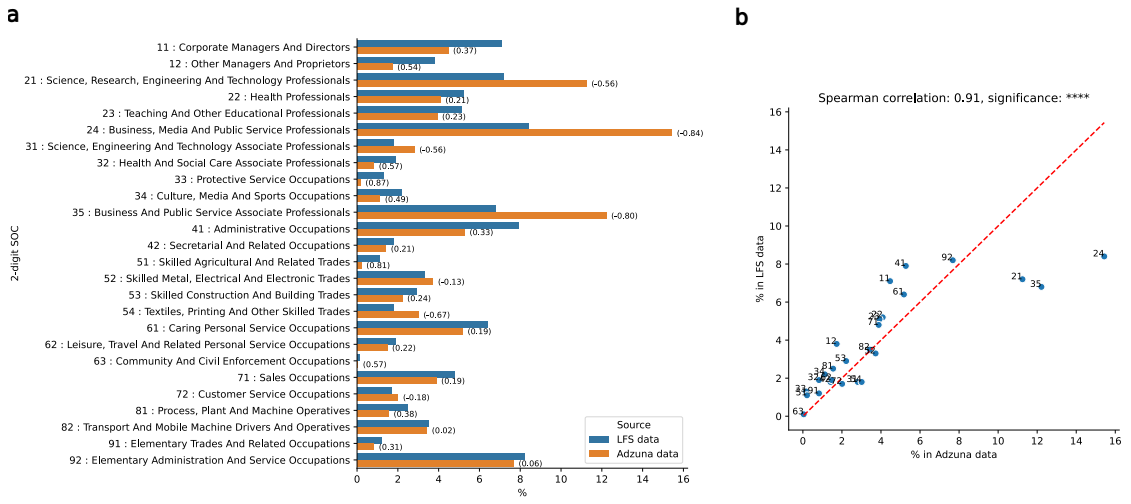

Fig 5: (a) Comparison between the percentage of adverts for each two-digit SOC occupation code in the Adzuna advert data set and the percentage of workers employed in each two-digit SOC occupation in the 2021 Labour Force Survey (LFS) [8]. Broad similarity is observed across both data sets, with some differences: Science, Research, Engineering and Technology Professionals, Business Media and Public Service Professionals, and Business and Public Service Associate Professionals are over-represented in the Adzuna data set, whereas Skilled Agricultural and Related Trades and Protective Service Occupations are under-represented in the job adverts data set. Under-representation of senior management roles in the adverts data sets is also apparent. (b) Scatter plot of the percentage of adverts in the Adzuna data set and the percentage of workers in the 2021 LFS for each two-digit SOC occupation. Each point is annotated with its two-digit SOC code, as in (a).

C.2 Comparison of Adzuna predicted salaries and official wage statistics

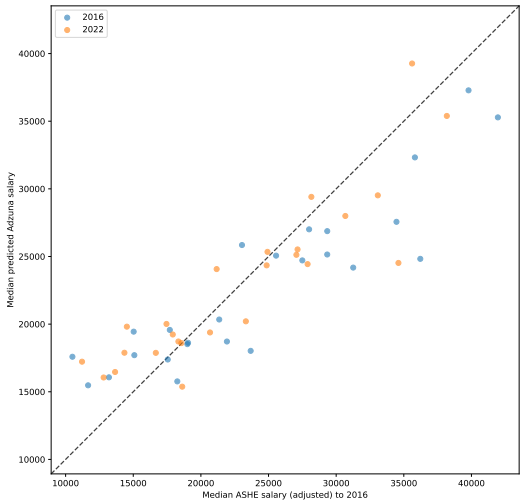

Fig 6: Comparison between median salary adjusted for inflation to 2016 prices from the Annual Survey of Hours and Earning (ASHE) and unadjusted Adzuna predicted salary for each two-digit SOC occupation code [9]. Broad agreement is observed between salaries in the Adzuna job advert data set and ASHE survey data across both 2016 and 2022. Note that Adzuna has higher predicted salaries than ASHE for lower paid occupations, and *vice versa* for higher paid occupations.

C.3 The geographic variability of salaries and power power purchasing parity

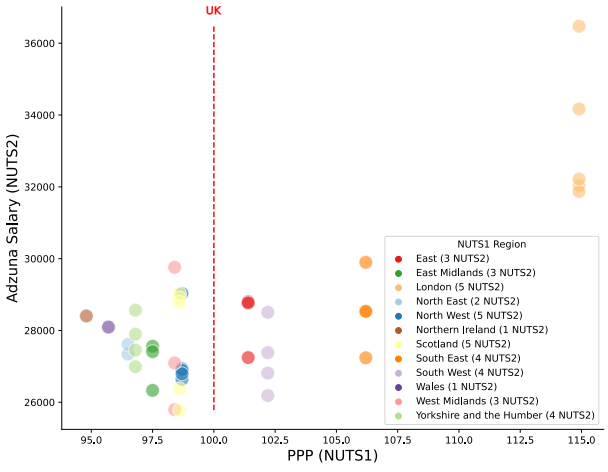

Fig 7: Comparison of average salary for Adzuna job adverts from each NUTS2 region against the estimated purchasing power parity (PPP) of their corresponding NUTS1 region as derived by [10] relative to UK average (100). We find small variations in salary across most regions (although their PPP varies from 95 to 106) and only NUTS2 regions in London have both larger average salary for advertised jobs and higher PPP.

References

[1] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *Proceedings of the national academy of sciences*, 107(29):12755–12760, 2010.

[2] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[3] Jean-Charles Delvenne, Michael T Schaub, Sophia N Yaliraki, and Mauricio Barahona. The stability of a graph partition: A dynamics-based framework for community detection. *Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems*, pages 221–242, 2013.

[4] Zijing Liu and Mauricio Barahona. Graph-based data clustering via multiscale community detection. *Applied Network Science*, 5:1–20, 2020.

[5] Robert G Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[6] Renaud Lambiotte, Jean-Charles Delvenne, and Mauricio Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, 2014.

[7] Michael Scheutzow and Dominik Schindler. Convergence of Markov chain transition probabilities. *Electronic Communications in Probability*, 26:1 – 13, 2021.

[8] Labour force survey - office for national statistics. Available from: https://www.nomisweb.co.uk/sources/aps.

[9] Office for National Statistics. Annual survey of hours and earnings. Available from: https://www.ons.gov.uk/surveys/informationforbusinesses/businesssurveys/annualsurveyofhoursandearningsashe.

[10] David Hearne and Alex de Ruyter. *Regional success after Brexit: The need for new measures*. Brexit Studies Series. Emerald Publishing Limited, 2019.