Understanding within-occupation heterogeneity in skillsets using large online job vacancy data

Shruti Sarika Chakraborty

University of Oxford, Computer Science

Tytus Wilam

University of Chicago, Sociology

Motivation

- Shocks like technology affects what skills are needed in the labor force,
- Workers are displaced and occupations transformed,
- Negative political externalities compound other problems, including catastrophic risks.

- Design appropriate education, skills, and industrial policies for a labor market that has a changing demand for skills,
- There exists a large literature interested in the impact of AI on work,
- This project questions how ai-exposure is currently calculated.

Current state-of-the-art relies on national surveys (example of ai and ml)

1. Assume tasks are known from surveys, (!)
2. Construct a measure of each task’s exposure to automation,
3. Construct an ai-exposure score for each occupation as a weighted average of the ai-exposure of its component tasks,
4. Construct firm- / industry- / area-level ai-exposure scores as a weighted average of the ai-exposure of its component tasks,
5. Use those occupation/firm/industry/area measures to explain wage movements, hiring decisions, and worker trajectories.

We use online job vacancies instead of surveys

Surveys (O*NET)	Online Job Vacancies (Lightcast)
statistically valid for their purposes with 20k yearly respondents and with good understanding of blind spots, but costly and with low and falling response rate (most recent wave at 39%),	full population of online vacancies (~20,000,000 job postings for 2019),
forced homogeneity of occupations (although in principle possible to update or merge occupations),	allows for heterogeneity of skillsets within occupation,
captures average members of an occupation at the national level,	captures variation in the local labor markets and is, in principle, not limited to the US,
static data,	real time data,
stock measure not a flow measure,	flow measure,
data collected for the analysis of tasks.	data originally created for a different purpose.

Outstanding questions

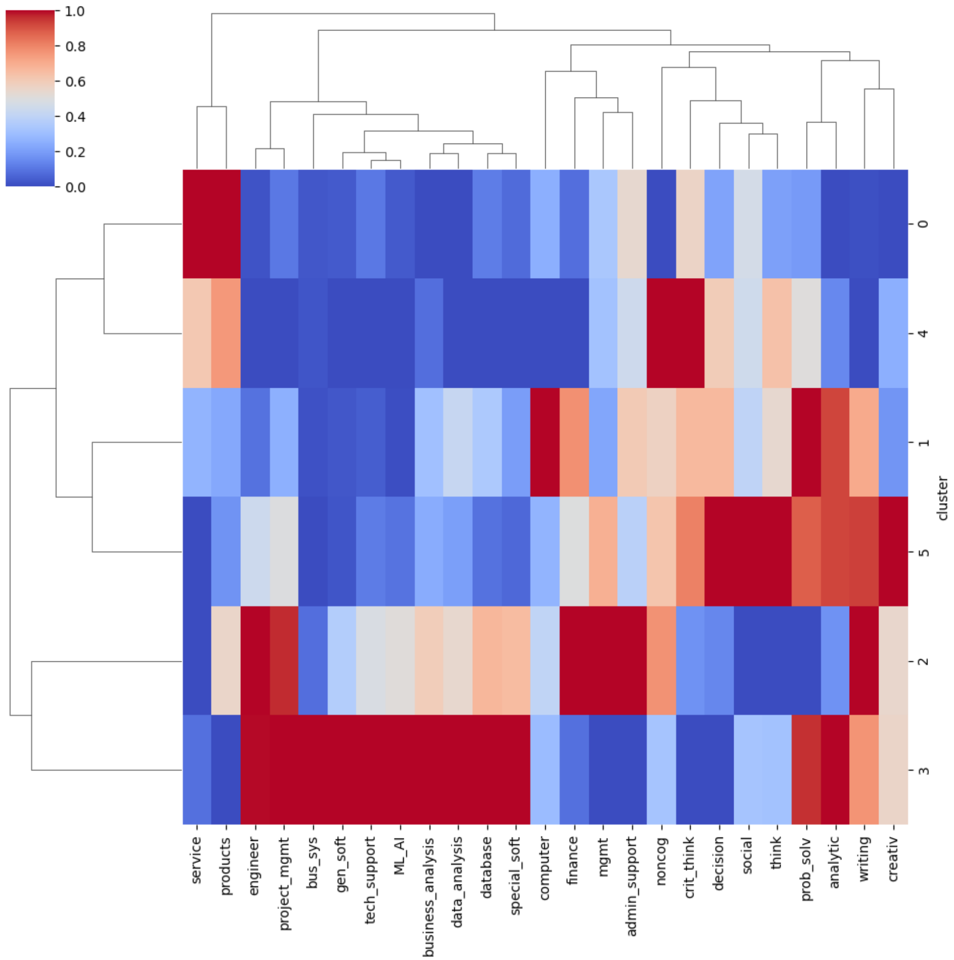
- **Regional and sectoral disparate impact:** Occupations contain persons with different skillsets who might be differentially affected by technology.
- **Occupational change in the face of technological change:** doctors and lawyers who code vs. hospitals that hire data scientists or outsource,
- **Control over new tasks:** Who will do the new tasks?

Data and method

- 20 million observations for one year, 2019 (we use 300k for tests)
- From Lightcast, a private provider
- Occupation, metropolitan area, industry, and 25 skills described for each observation, including 8 digital skills observation
- Try on a small sample:
    - Keep only popular occupations (n > 500) → 250k observations,
    - Formed skill clusters with different methods and parameters (k-means with different number of clusters, DBSCANS, hierarchical clustering)
    - Form skill clusters clusters for technical skills,
    - Choose the optimal clustering method (elbow method, Davies-Bouldin, Calinski-Harabasz),
    - Each cluster gives us a unique combination of skills,
    - Each occupation contains different skillsets.
- Repeat for all observations

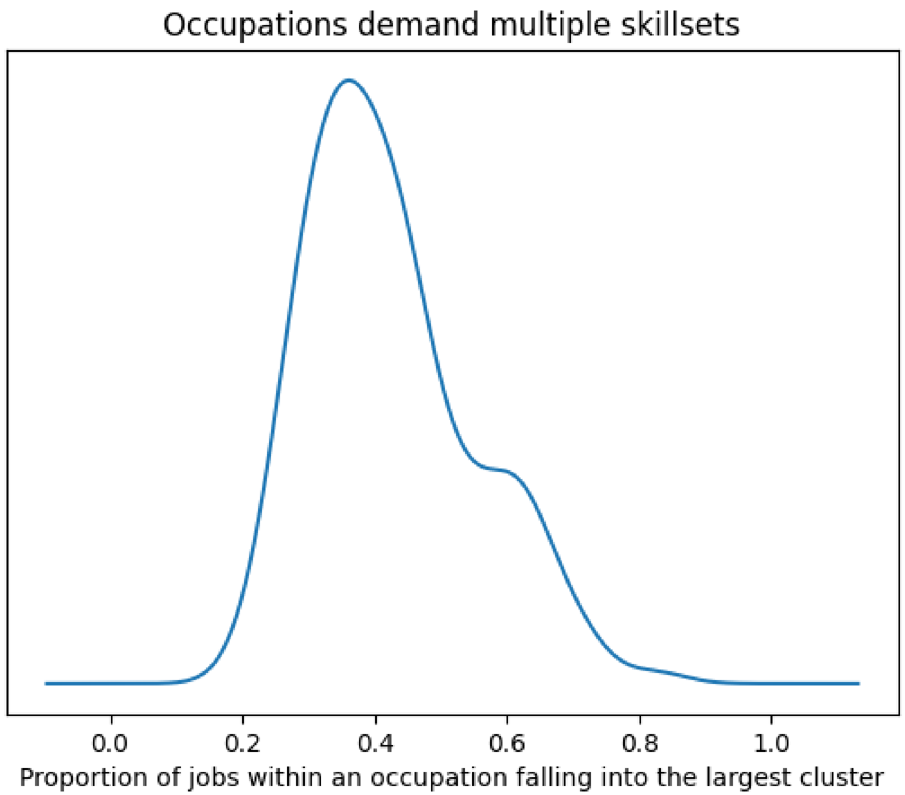
Association of the skills in each cluster for cluster type 1 (25 skills taken together, forming 6 clusters)

CLUSTER 0		CLUSTER 1		CLUSTER 2		CLUSTER 3		CLUSTER 4		CLUSTER 5	
products	0.17756	service	0.418681	social	0.11901	special_soft	0.15715	noncog	0.18754	social	0.30329
social	0.15813	noncog	0.106371	noncog	0.11877	social	0.10489	admin_support	0.13644	service	0.15648
service	0.14359	admin_support	0.090448	computer	0.11615	project_mgmt	0.07932	project_mgmt	0.08678	noncog	0.13384
noncog	0.11145	products	0.084207	admin_support	0.07746	tech_support	0.06028	products	0.07335	admin_support	0.061
admin_support	0.05167	computer	0.048732	service	0.0658	noncog	0.05987	mgmt	0.07192	computer	0.04427
computer	0.04996	mgmt	0.039647	project_mgmt	0.06317	gen_soft	0.05905	computer	0.07076	mgmt	0.04364
mgmt	0.04419	project_mgmt	0.033318	finance	0.06281	bus_sys	0.05216	finance	0.05997	project_mgmt	0.04106
bus_sys	0.03483	finance	0.028659	special_soft	0.04908	database	0.04537	special_soft	0.0583	prob_solv	0.03617
finance	0.03408	special_soft	0.022706	prob_solv	0.0474	writing	0.03807	writing	0.04536	writing	0.0323
prob_solv	0.03335	writing	0.019453	mgmt	0.04669	ML_AI	0.0376	prob_solv	0.02511	finance	0.02592
project_mgmt	0.02647	prob_solv	0.019276	writing	0.04395	prob_solv	0.03613	bus_sys	0.0242	bus_sys	0.01822
writing	0.02518	bus_sys	0.015425	bus_sys	0.03003	computer	0.03284	database	0.02413	creativ	0.0173
creativ	0.02338	database	0.014263	products	0.02221	data_analysis	0.03021	tech_support	0.02228	database	0.01122
special_soft	0.01844	tech_support	0.010413	business_anal	0.02026	service	0.02806	creativ	0.02073	tech_support	0.01
business_analysis	0.01484	creativ	0.009494	database	0.01892	mgmt	0.02779	engineer	0.02021	think	0.00965
tech_support	0.00868	business_analysis	0.007192	data_analysis	0.01802	products	0.0266	business_anal	0.01999	crit_think	0.00847
decision	0.00692	think	0.006385	creativ	0.01662	business_anal	0.02488	data_analysis	0.01425	business_anal	0.00812
database	0.00653	crit_think	0.006053	tech_support	0.01318	engineer	0.02373	ML_AI	0.00815	decision	0.00789
analytic	0.00624	data_analysis	0.005488	analytic	0.01102	finance	0.02118	gen_soft	0.00814	special_soft	0.00731
data_analysis	0.00541	gen_soft	0.004836	think	0.00899	creativ	0.01987	think	0.0064	data_analysis	0.00701
gen_soft	0.00518	decision	0.003209	decision	0.00784	admin_support	0.01331	decision	0.00613	analytic	0.00607
think	0.00515	engineer	0.002457	crit_think	0.00706	analytic	0.00848	crit_think	0.00562	engineer	0.0051
engineer	0.00366	ML_AI	0.00187	engineer	0.00676	think	0.00546	analytic	0.00426	gen_soft	0.00373
crit_think	0.00341	analytic	0.001416	gen_soft	0.00676	decision	0.00406	service	0	ML_AI	0.00195
ML_AI	0.00171	social	0	ML_AI	0.00208	crit_think	0.00365	social	0	products	0



Six general skillsets, three technical skillsets

151132	**0**	406
	1	217
	2	1891
	3	5262
	4	185
	5	537
151134	**0**	42
	1	28
	2	375
	3	814
	4	53
	5	147
151141	**0**	51
	1	63
	2	245
	3	743
	4	26
	5	56



Multiple skillsets per occupation

Possible next steps?

- Construct the clusters for the full population,
- **Examine the distribution of skillsets by occupation / firm / sector / region**,
  - Are there skillsets that fall under different occupations in different regions?
- Examine if within-occupation wage inequality is related to within-occupation task polarization,
- Examine patterns of integration of new technologies into new occupations and sectors through case studies. When is there a need for technical expertise? If there is such a need, who does it (occupation and location)?
- Show that the change in skills precedes the change in the O*NET categorization scheme.

Main drawbacks of our approach:

- No direct wage measurement,
- Only coarse skill measurement,
- Flow, not stock data,
- Skills mean different things in different occupations,
- Occupation could turn out to have a strong effect net of skill,
- We use a different measurement system from the one developed by O*NET which makes it difficult to link to existing literature on tasks.