Are occupations "bundles of skills"? Identifying latent skill profiles in the labour market using topic modeling

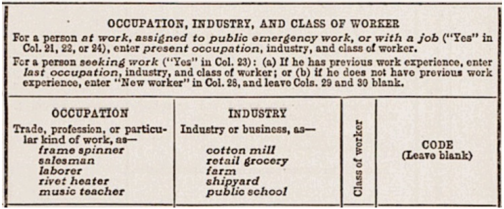
ODISSEI Conference
2 November 2023

Marie Labussière & Thijs Bol
University of Amsterdam

European Research Council

career

Introduction



OCCUPATION, INDUSTRY, AND CLASS OF WORKER

For a person *at work, assigned to public emergency work, or with a job* ("Yes" in Col. 21, 22, or 24), enter *present occupation*, industry, and class of worker.

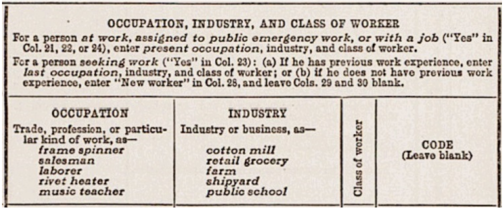
For a person *seeking work* ("Yes" in Col. 23): (a) If he has previous work experience, enter *last occupation*, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank.

OCCUPATION	INDUSTRY	Class of worker	CODE (Leave blank)
Trade, profession, or particular kind of work, as— *frame spinner* *salesman* *laborer* *rivet heater* *music teacher*	Industry or business, as— *cotton mill* *retail grocery* *farm* *shipyard* *public school*		

1 proprietors
2 clerical employees
3 skilled workers
4 laborers

Bureau of the US Census, 1897 (above) and 1940 (left)

Introduction


OCCUPATION, INDUSTRY, AND CLASS OF WORKER

For a person *at work, assigned to public emergency work, or with a job* ("Yes" in Col. 21, 22, or 24), enter *present occupation*, industry, and class of worker.

For a person *seeking work* ("Yes" in Col. 23): (a) If he has previous work experience, enter *last occupation*, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank.

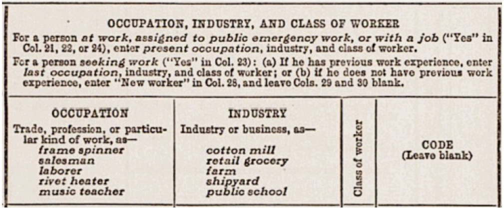
OCCUPATION	INDUSTRY	Class of worker	CODE (Leave blank)
Trade, profession, or particular kind of work, as— *frame spinner* *salesman* *laborer* *rivet heater* *music teacher*	Industry or business, as— *cotton mill* *retail grocery* *farm* *shipyard* *public school*		

```
1    proprietors
2    clerical employees
3    skilled workers
4    laborers
```

Bureau of the US Census, 1897 (above) and 1940 (left)

- A central unit for understanding labor market inequalities

Introduction



OCCUPATION, INDUSTRY, AND CLASS OF WORKER

For a person *at work, assigned to public emergency work, or with a job* ("Yes" in Col. 21, 22, or 24), enter *present occupation*, industry, and class of worker.

For a person *seeking work* ("Yes" in Col. 23): (a) If he has previous work experience, enter *last occupation*, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank.

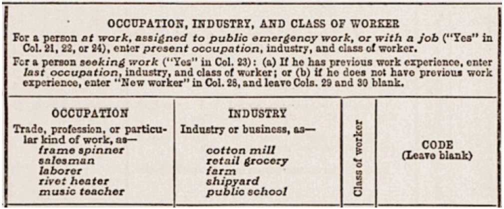
OCCUPATION	INDUSTRY	Class of worker	CODE (Leave blank)
Trade, profession, or particular kind of work, as— *frame spinner* *salesman* *laborer* *rivet heater* *music teacher*	Industry or business, as— *cotton mill* *retail grocery* *farm* *shipyard* *public school*		

```
1   proprietors
2   clerical employees
3   skilled workers
4   laborers
```

Bureau of the US Census, 1897 (above) and 1940 (left)

- A central unit for understanding labor market inequalities
- Why are occupations so important?

Introduction


OCCUPATION, INDUSTRY, AND CLASS OF WORKER

For a person *at work, assigned to public emergency work, or with a job* ("Yes" in Col. 21, 22, or 24), enter *present occupation*, industry, and class of worker.

For a person *seeking work* ("Yes" in Col. 23): (a) If he has previous work experience, enter *last occupation*, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank.

OCCUPATION	INDUSTRY	Class of worker	CODE (Leave blank)
Trade, profession, or particular kind of work, as— *frame spinner*, *salesman*, *laborer*, *rivet heater*, *music teacher*	Industry or business, as— *cotton mill*, *retail grocery*, *farm*, *shipyard*, *public school*		

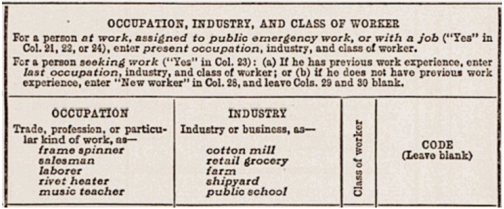
```
1   proprietors
2   clerical employees
3   skilled workers
4   laborers
```

Bureau of the US Census, 1897 (above) and 1940 (left)

- A central unit for understanding labor market inequalities
- Why are occupations so important?

  $\rightarrow$ the nature of skills and tasks performed at work

Introduction



OCCUPATION, INDUSTRY, AND CLASS OF WORKER

For a person *at work, assigned to public emergency work, or with a job* ("Yes" in Col. 21, 22, or 24), enter *present occupation*, industry, and class of worker.

For a person *seeking work* ("Yes" in Col. 23): (a) If he has previous work experience, enter *last occupation*, industry, and class of worker; or (b) if he does not have previous work experience, enter "New worker" in Col. 28, and leave Cols. 29 and 30 blank.

OCCUPATION	INDUSTRY	Class of worker	CODE (Leave blank)
Trade, profession, or particular kind of work, as— *frame spinner* *salesman* *laborer* *rivet heater* *music teacher*	Industry or business, as— *cotton mill* *retail grocery* *farm* *shipyard* *public school*		

```
1   proprietors
2   clerical employees
3   skilled workers
4   laborers
```

Bureau of the US Census, 1897 (above) and 1940 (left)

- A central unit for understanding labor market inequalities
- Why are occupations so important?
  - → the nature of skills and tasks performed at work
  - → occupations are coherent *bundles* of skills (Acemoglu and Autor, 2011; Mouw and Kalleberg, 2010)

Research question

Questioned by empirical evidence:

- heterogeneity in task content (Yamaguchi, 2012; Autor and Handel, 2013; Freeman et al., 2020)

Research question

Questioned by empirical evidence:

- heterogeneity in task content (Yamaguchi, 2012; Autor and Handel, 2013; Freeman et al., 2020)
- significant overlap in workers' skill portfolios (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Cheng and Park, 2020)

Research question

Questioned by empirical evidence:

- heterogeneity in task content (Yamaguchi, 2012; Autor and Handel, 2013; Freeman et al., 2020)
- significant overlap in workers' skill portfolios (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Cheng and Park, 2020)

Are occupations made up of well-defined and homogeneous "bundles of skills"?

Research question

Questioned by empirical evidence:

- heterogeneity in task content (Yamaguchi, 2012; Autor and Handel, 2013; Freeman et al., 2020)
- significant overlap in workers' skill portfolios (Poletaev and Robinson, 2008; Gathmann and Schönberg, 2010; Cheng and Park, 2020)

Are occupations made up of well-defined and homogeneous "bundles of skills"?

Case: the UK

Theory

From skills to skill profiles

Skills are often poorly conceptualized (Liu and Grusky, 2013)

- Skills are analyzed in isolation

Theory

From skills to skill profiles

Skills are often poorly conceptualized (Liu and Grusky, 2013)

- Skills are analyzed in isolation

  $\rightarrow$ context matters

Theory

From skills to skill profiles

Skills are often poorly conceptualized (Liu and Grusky, 2013)

- Skills are analyzed in isolation

  $\rightarrow$ context matters

  $\rightarrow$ **skill profiles**: mixes of (different types of) hard and soft skills

Theory

From skills to skill profiles

Skills are often poorly conceptualized (Liu and Grusky, 2013)

- Skills are analyzed in isolation

  → context matters

  → **skill profiles**: mixes of (different types of) hard and soft skills

Theory

From skills to skill profiles

Skills are often poorly conceptualized (Liu and Grusky, 2013)

- Skills are analyzed in isolation

  $\rightarrow$ context matters

  $\rightarrow$ **skill profiles**: mixes of (different types of) hard and soft skills

$\rightarrow$ **Are occupations made of similar *skill profiles*?**

Data

Skills defined by employers at the job level

Data

Skills defined by employers at the job level

Online job ads gathered by the BGT in the UK:

Observation period	2019 calendar year
Full sample	6.9 million
Random Sample	600,000 (stratified by 2-digit occupation)

Occupation classification

Standard Occupational Classification 2010 (UK)

9 Major groups

25 Sub-Major groups

90 Minor groups

369 Unit groups

Example: postal workers

9	ELEMENTARY OCCUPATIONS
92	ELEMENTARY ADMINISTRATION AND SERVICE
921	ELEMENTARY ADMINISTRATION
9211	POSTAL WORKERS, MAIL SORTERS

Operationalisation

How to identify the skill profiles of occupations?

Job ads are already pre-processed by the BGT: extracts skills requirements and standardizes their labels

- 9,065 distinct skill requirements
- How to measure their association within the ads?

Operationalisation

How to identify the skill profiles of occupations?

Job ads are already pre-processed by the BGT: extracts skills requirements and standardizes their labels

- 9,065 distinct skill requirements
- How to measure their association within the ads?

  ➥ **topic modeling with LDA**

Method

Method: LDA Biterm topic model (Yan et al., 2013)

Method

Method: LDA Biterm topic model (Yan et al., 2013)

K=19 topics

Method

Method: LDA Biterm topic model (Yan et al., 2013)

K=19 topics

LDA outputs

$$topic_i = (p_{skill_1}, \ldots p_{skill_N}),\ \forall i{=}1,\ldots K$$

topic = latent skill category
$\hookrightarrow$ data-driven classification of skills

$$job_j = (p_{topic_1}, \ldots p_{topic_k}),\ \forall j{=}1,\ldots J$$

Method

Method: LDA Biterm topic model (Yan et al., 2013)

K=19 topics

LDA outputs

$$topic_i = (p_{skill_1}, \dots p_{skill_N}), \ \forall i{=}1,\dots K$$

topic = latent skill category
$\hookrightarrow$ data-driven classification of skills

$$job_j = (p_{topic_1}, \dots p_{topic_k}), \ \forall j{=}1,\dots J$$

$\hookrightarrow$ skill profiles of job postings

Results Biterm

19 latent topics or skill categories

Results Biterm

19 latent topics or skill categories

Skill category	**Skills, in decreasing order of probability**
Digital Marketing & content strategy	social media; marketing; digital marketing; creativity; marketing management; Google Analytics; market strategy; content management; copy writing; editing...

Results Biterm

19 latent topics or skill categories

Skill category	**Skills, in decreasing order of probability**
Digital Marketing & content strategy	social media; marketing; digital marketing; creativity; marketing management; Google Analytics; market strategy; content management; copy writing; editing...

Other skill categories

Project Management
Office administration & management
Sales & Business Development
Communication & Interpersonal Abilities
Caregiving & Support Services
Customer Service & Retail Operations
Financial operations
Web Development & Software Engineering
Logistics & Supply Chain Management
Engineering & Technical Expertise
Manufacturing & Engineering
Data Management & Analysis
Facility Maintenance
Healthcare & Patient Care
Business stragegy
Technical Support & Troubleshooting
Graphic Design & Creative Media
Scientific Research & Laboratory Work

Results Biterm

Job postings as skill profiles

Social Media Account Executive

Time Management, Content Management, Creative Writing, Social Media, Creativity, Business-to-Business, Social Media Platforms.

Results Biterm

Job postings as skill profiles

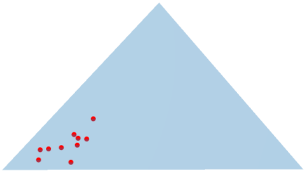
Social Media Account Executive

Time Management, Content Management, Creative Writing, Social Media, Creativity, Business-to-Business, Social Media Platforms.

Skill category	Probability
Digital Marketing and Content Strategy	0.75
Graphic Design and Creative Media	0.08
Communication and Interpersonal Abilities	0.07
Other skill categories	<0.05

Similarity of skill profiles between occupations

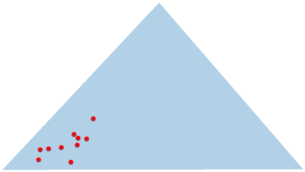
The topic space with three topics



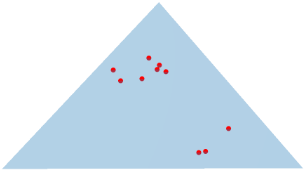
Documents mainly focus on one topic

Similarity of skill profiles between occupations

The topic space with three topics



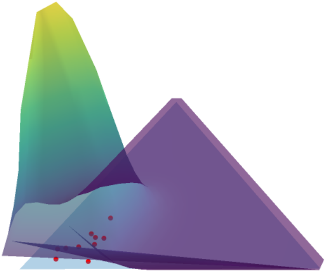
Documents mainly focus on one topic



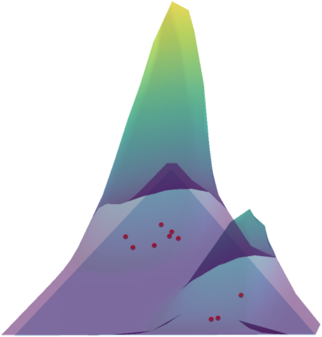
Documents mainly focus on two topics

Similarity of skill profiles between occupations

Empirical distributions with three topics



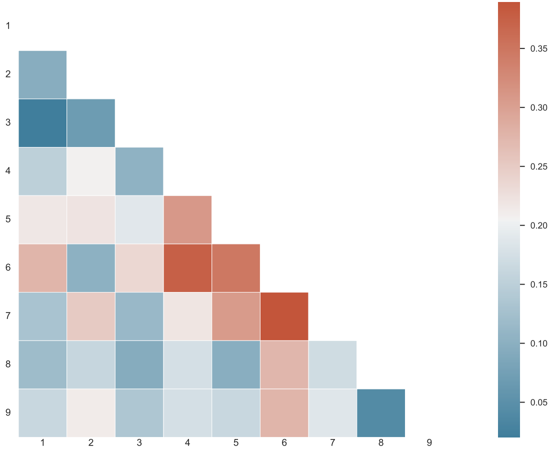
Documents mainly focus on one topic



Documents mainly focus on two topics

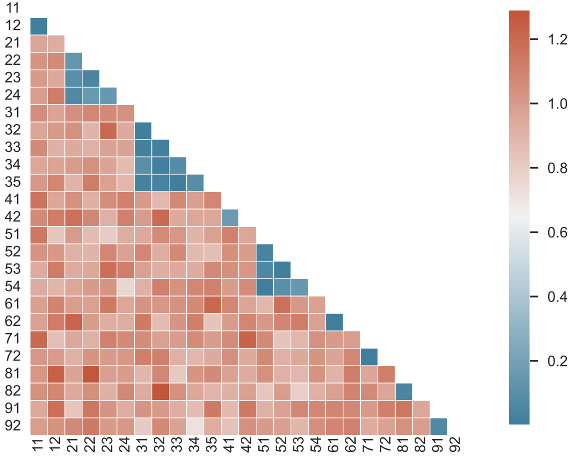
MMD distance between occupations

SOC Major groups



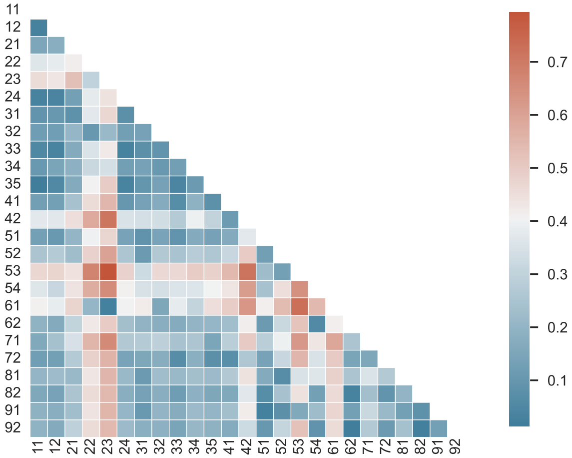
MMD distance between occupations

SOC Sub-Major groups (theoretical)



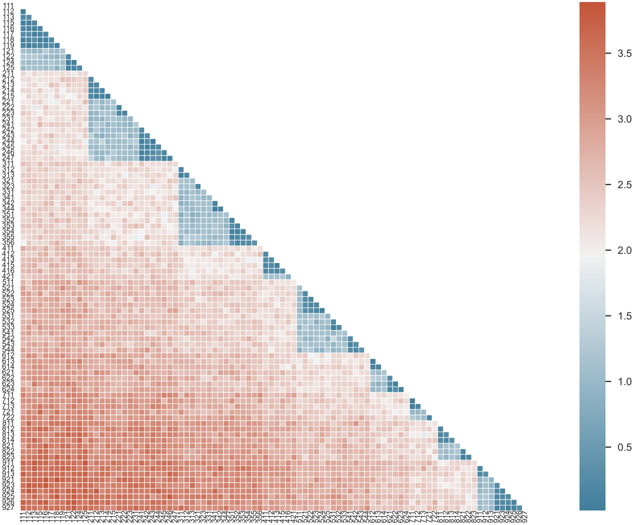
MMD distance between occupations

SOC Sub-Major groups (observed)



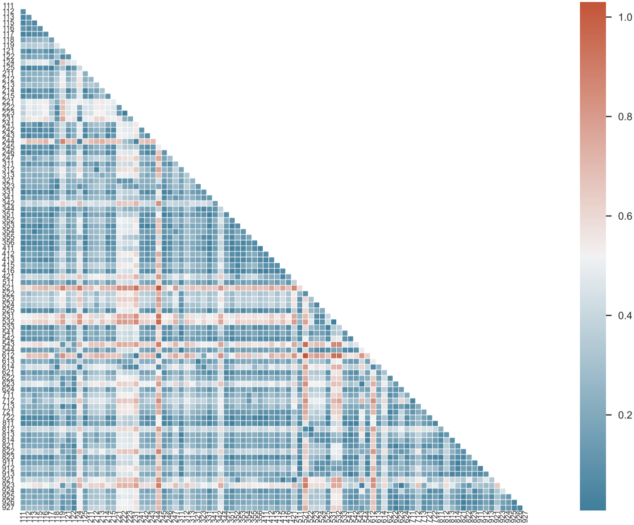
MMD distance between occupations

SOC Minor groups (theoretical)



MMD distance between occupations

SOC Minor groups (observed)

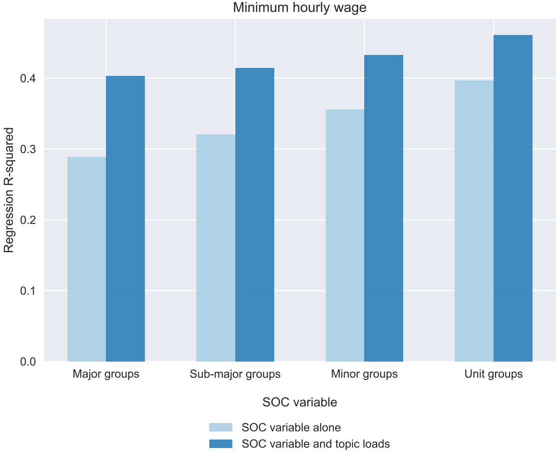


Wage analysis

Do the skill profiles capture substantive differences in job content?

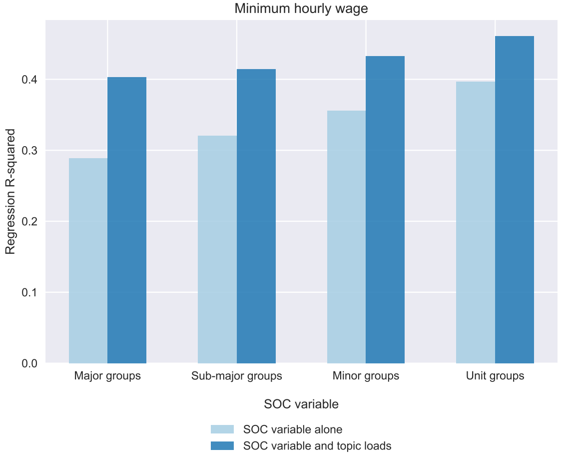
Wage analysis

Do the skill profiles capture substantive differences in job content?



Wage analysis

Do the skill profiles capture substantive differences in job content?



➥ **Occupation and skill profile are complementary.**

Conclusion

1. Occupations are not good proxy for job skill content (Freeman et al., 2020; Poletaev and Robinson, 2008)
2. Proximity of skill profiles: more room for mobility than usually assumed? (DeMaria et al., 2020)
3. Occupations and skill profiles bring complementary information

Conclusion

1. Occupations are not good proxy for job skill content (Freeman et al., 2020; Poletaev and Robinson, 2008)
2. Proximity of skill profiles: more room for mobility than usually assumed? (DeMaria et al., 2020)
3. Occupations and skill profiles bring complementary information

⏭ What's in an occupation? ⏮

Discussion

Want to know more?

projectcareer.eu

@CAREER_erc

References I

Acemoglu, D. and Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., and Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science Advances*, 4(7):eaao6030.

Autor, D. H. and Handel, M. J. (2013). Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics*, 31(S1):S59–S96.

Cheng, S. and Park, B. (2020). Flows and Boundaries: A Network Approach to Studying Occupational Mobility in the Labor Market. *American Journal of Sociology*, 126(3):577–631.

DeMaria, K., Fee, K., and Wardrip, K. (2020). Exploring A Skills-Based Approach To Occupational Mobility. Community Affairs Discussion Paper 89004, Federal Reserve Bank of Philadelphia.

Djumalieva, J. and Sleeman, C. (2018). An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. In Larsen, C., Rand, S., Schmid, A., and Dean, A., editors, *Developing Skills in a Changing World of Work*, pages 425–454. Rainer Hampp Verlag.

Freeman, R. B., Ganguli, I., and Handel, M. J. (2020). Within-Occupation Changes Dominate Changes in What Workers Do: A Shift-Share Decomposition, 2005–2015. *AEA Papers and Proceedings*, 110:394–399.

Gathmann, C. and Schönberg, U. (2010). How General Is Human Capital? A Task-Based Approach. *Journal of Labor Economics*, 28(1):1–49.

Liu, Y. and Grusky, D. B. (2013). The Payoff to Skill in the Third Industrial Revolution. *American Journal of Sociology*, 118(5):1330–1374.

Mouw, T. and Kalleberg, A. L. (2010). Occupations and the Structure of Wage Inequality in the United States, 1980s to 2000s. *American Sociological Review*, 75(3):402–431.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

References II

Poletaev, M. and Robinson, C. (2008). Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984–2000. *Journal of Labor Economics*, 26(3):387–420.

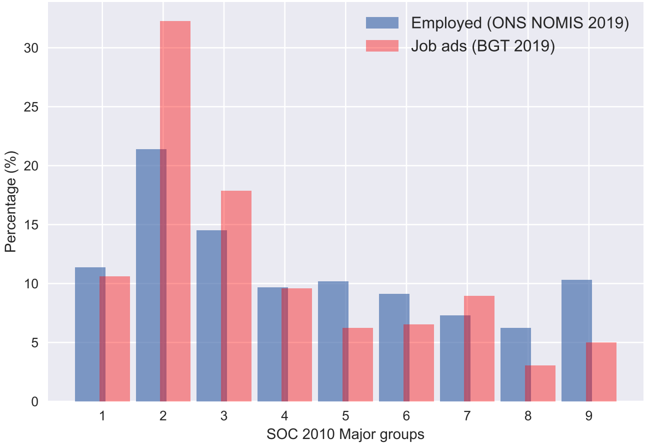
Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Yamaguchi, S. (2012). Tasks and Heterogeneous Human Capital. *Journal of Labor Economics*, 30(1):1–53.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 1445–1456, Rio de Janeiro, Brazil. ACM Press.

Data

Representativity



Theory

An unquestioned assumption about the nature of occupations

"Task models provide a natural framework for interpreting patterns related to occupations in the labor market, (...) since **we can think of occupations as bundles of tasks.**" (p.1118) (Acemoglu and Autor, 2011)

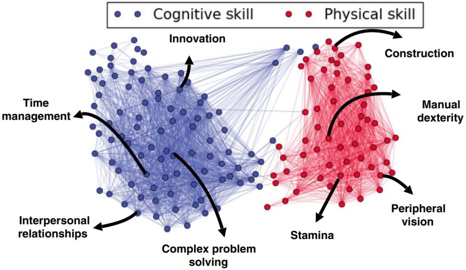
"First, occupations vary in their **skill**, that is, the **degree** of complexity of occupational activities and the **amount** of training time required to perform them adequately." (p.404) (Mouw and Kalleberg, 2010)

How to map the skill structure of the labour market?

Prevalence of top-down approaches

Exceptions Alabdulkareem et al. (2018); Djumalieva and Sleeman (2018) :

Identification of communities of *similar* skills



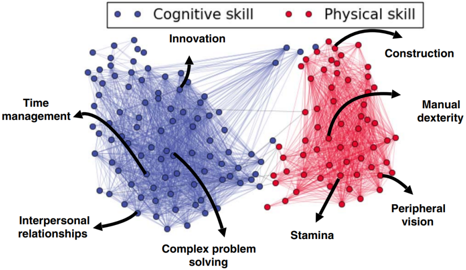
Alabdulkareem et al. (2018)

How to map the skill structure of the labour market?

Prevalence of top-down approaches

Exceptions Alabdulkareem et al. (2018); Djumalieva and Sleeman (2018) :

Identification of communities of *similar* skills



Alabdulkareem et al. (2018)

- ✗ one skill belongs to one and only one category
- ✗ it does not capture patterns of complementarity

Biterm Topic model (BTM)

A variant of Latent Dirichlet Allocation (LDA)

Job ads are short texts: **data sparsity** and **lack of context**

The biterm topic model (Yan et al., 2013)

- BTM uses biterms instead of words as semantic units

$$\text{“planning sales Excel”} = \begin{cases} \text{“planning sales”} \\ \text{“sales Excel”} \\ \text{“Excel planning”} \end{cases}$$

Biterm Topic model (BTM)

A variant of Latent Dirichlet Allocation (LDA)

Job ads are short texts: **data sparsity** and **lack of context**

The biterm topic model (Yan et al., 2013)

- BTM uses biterms instead of words as semantic units

$$\text{“planning sales Excel”} = \begin{cases} \text{“planning sales”} \\ \text{“sales Excel”} \\ \text{“Excel planning”} \end{cases}$$

- It directly models the generation of words co-occurrence patterns in the whole corpus ($\neq$ in each single document)
- The document generative process can be estimated

Biterm topic model

From Yan et al. (2013)

Biterm: unordered word-pair co-occurrence in a short context

Generative process

- The corpus consists of a mixture of topics
- Each biterm is drawn from a specific topic

1. For each topic z, draw a topic-specific word distribution $\phi_z \backsim Dir(\beta)$
2. Draw a topic distribution $\theta \backsim Dir(\alpha)$ for the whole collection
3. For each biterm b in the biterm set B
   - draw a topic assignment $z \backsim Multi(\theta)$
   - draw two words: $w_i, w_j \backsim Multi(\phi_z)$

Biterm topic model

From Yan et al. (2013)

The joint probability of a biterm $b = (w_i, w_j)$:

$$P(b) = \sum_z P(z)P(w_i|z)P(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z}$$

The likelihood of the whole corpus:

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z}$$

The topic proportion of a document can be estimated via Bayes and the empirical distribution of the generated biterms:

$$P(z|d) = \sum_b P(z|b)P(b|d) = \frac{P(z)P(w_i|z)P(w_j|z)}{\sum_z P(z)P(w_i|z)P(w_j|z)} \frac{n_d b}{\sum_b n_d(b)}$$

Biterm topic model

Comparison with LDA (from Yan et al. (2013))

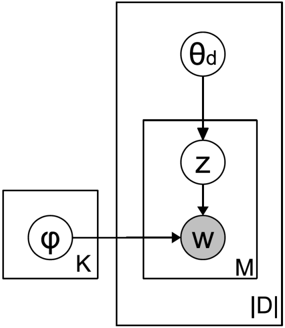


Plate notation LDA

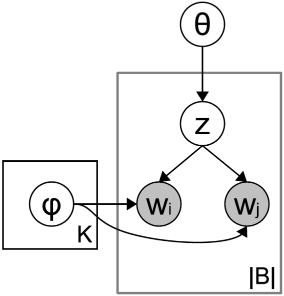


Plate notation Biterm

Setting for LDA

Biterm model with k topics

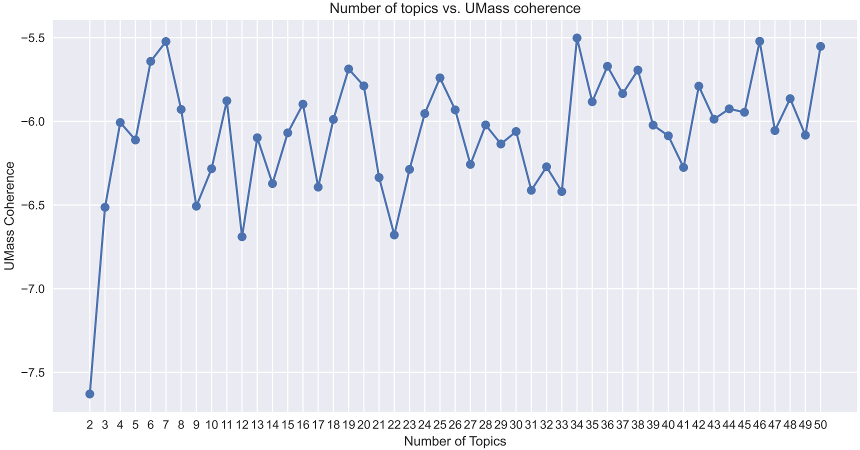
`maximtrp/bitermplus` (cythonized)

Priors	$\alpha = \beta = \frac{1}{k}$ i.e., the ads/topics are specialised
Iterations	2,000
Choice of k	✓ Visual inspection, Perplexity ✗ Coherence

Good compromise: $\boldsymbol{k} = 19$ topics/skill categories

Choice of the number of topics

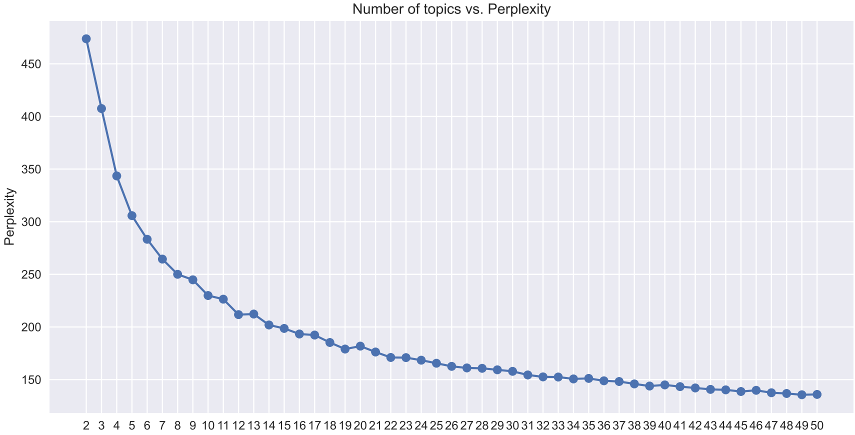
Coherence (Röder et al., 2015)



Back

Choice of the number of topics

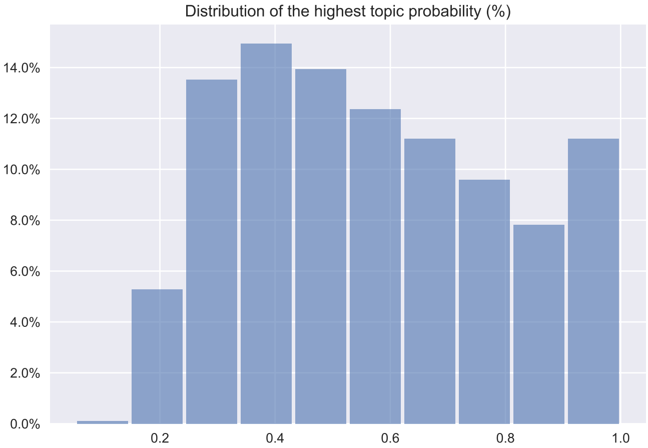
Perplexity



◂ Back

Distribution of job ads over the topics

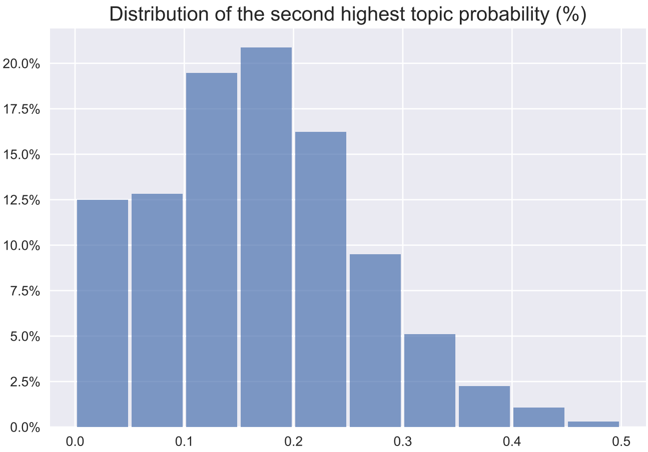
Highest topic probability



Back

Distribution of job ads over the topics

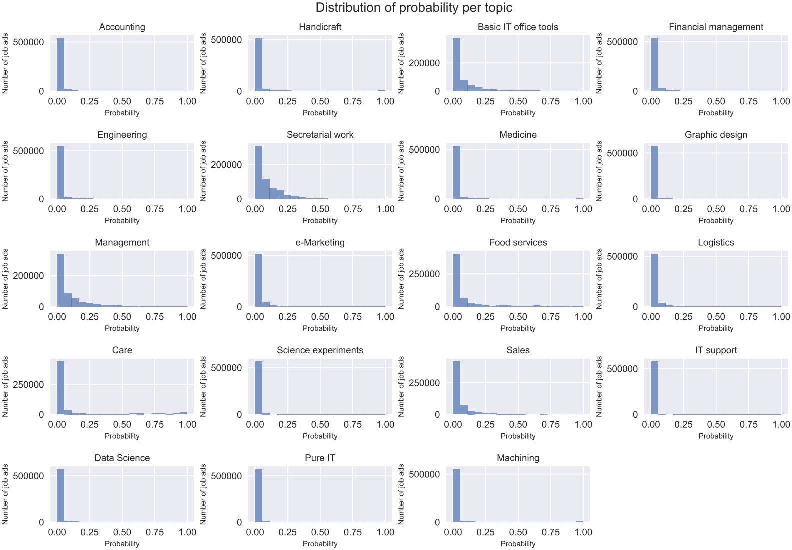
Second highest topic probability



Back

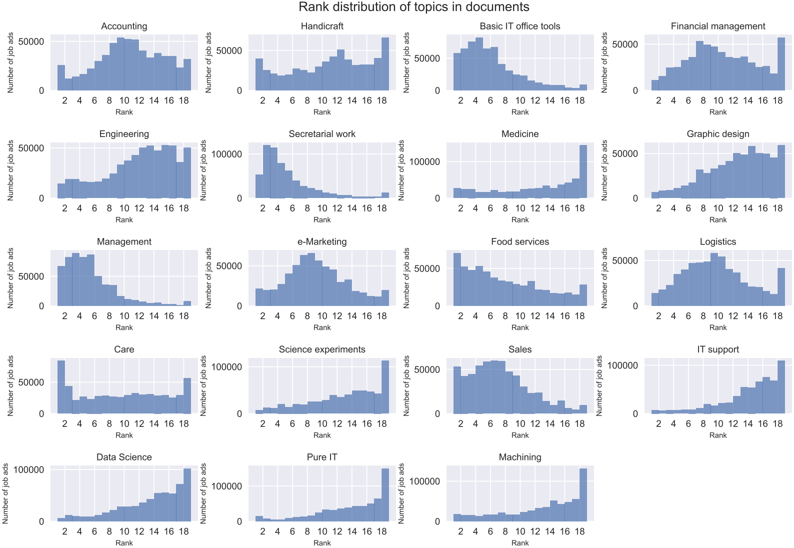
Distribution of job ads over the topics

Probability distribution



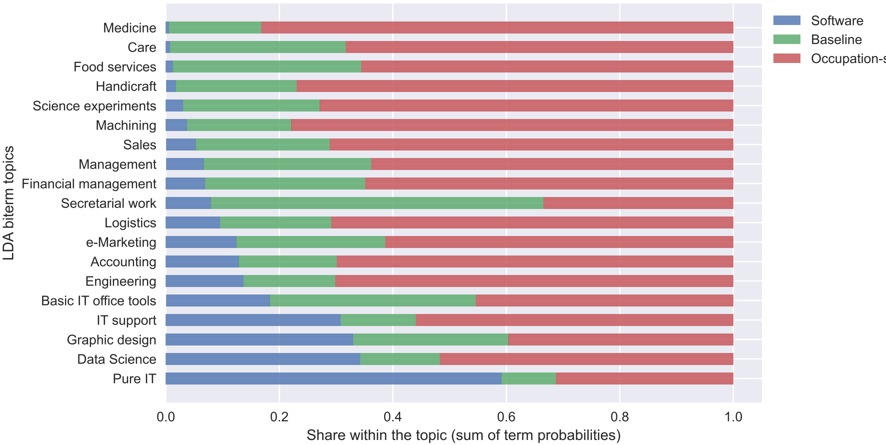
Distribution of job ads over the topics

Rank distribution



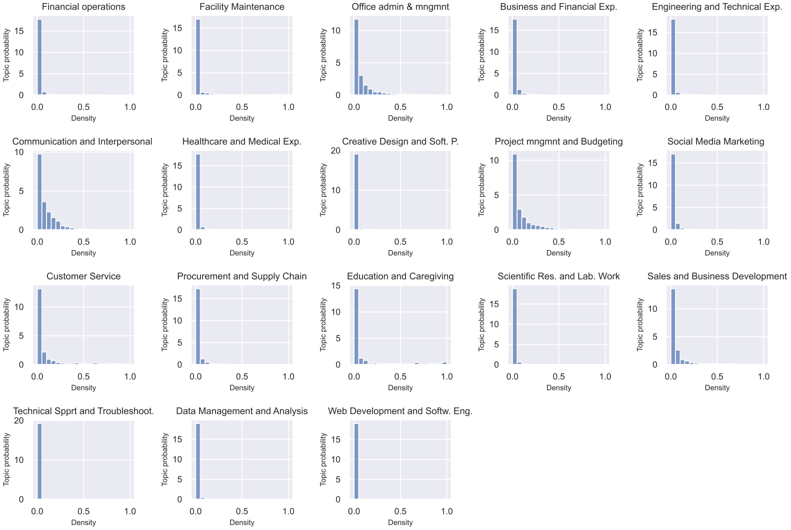
Type of skills within the skill sets

Share of soft and software skills



Marginal distributions

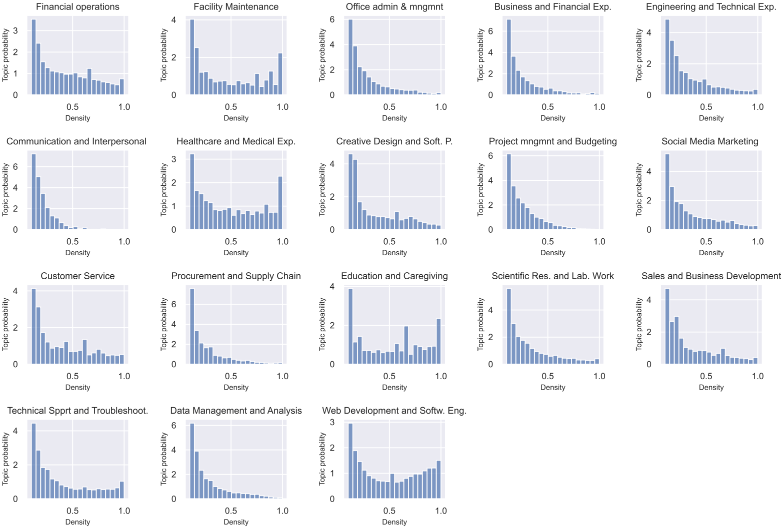
zero included



Back

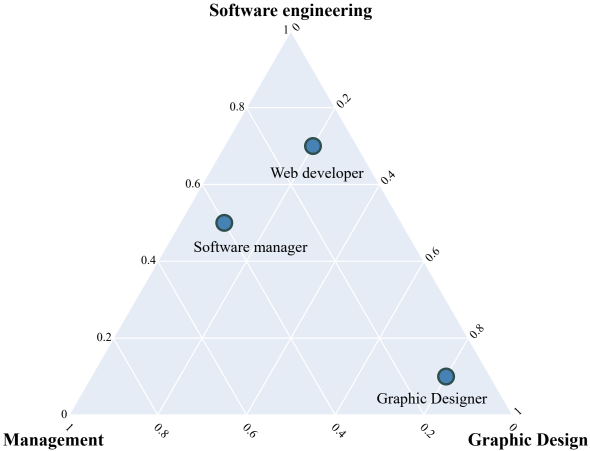
Marginal distributions

minimum set to 0.1



Back

Illustration 2-simplex



Similarity of skill profiles within/between occupations

Maximum mean discrepancy (MMD)

Strategy:

1. define an empirical distribution over the job ads at the occupation level
   $\hookrightarrow$ with the job ads defined by their 19-dimension vectors.
2. compare the empirical distributions of occupations: how much do they overlap/differ?

Similarity of skill profiles within/between occupations

Maximum mean discrepancy (MMD)

Strategy:

1. define an empirical distribution over the job ads at the occupation level
   $\hookrightarrow$ with the job ads defined by their 19-dimension vectors.
2. compare the empirical distributions of occupations: how much do they overlap/differ?

Maximum mean discrepancy

Kernel-based distance between probability distributions
✓ Non-parametric
✓ Implementable in high dimension
✓ Robust

KDE

Kernel density estimation

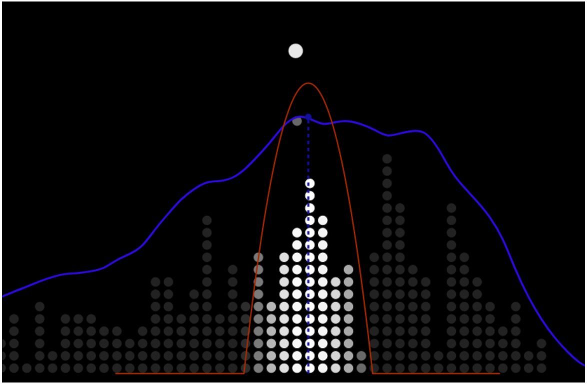


Image from a beautiful animation by Matthew Conlen
https://mathisonian.github.io/kde/

MMD

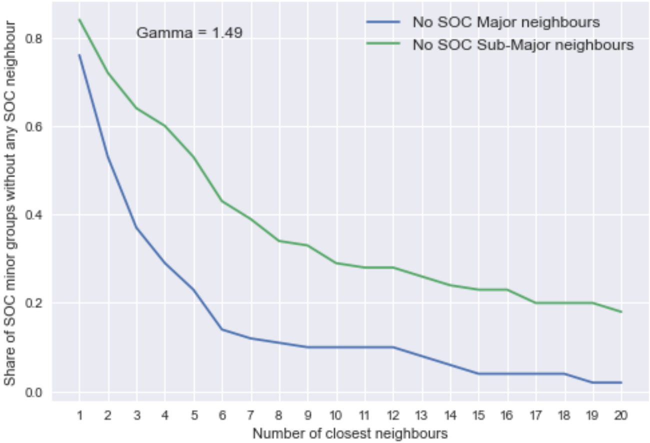
Maximum mean discrepancy

Technical details

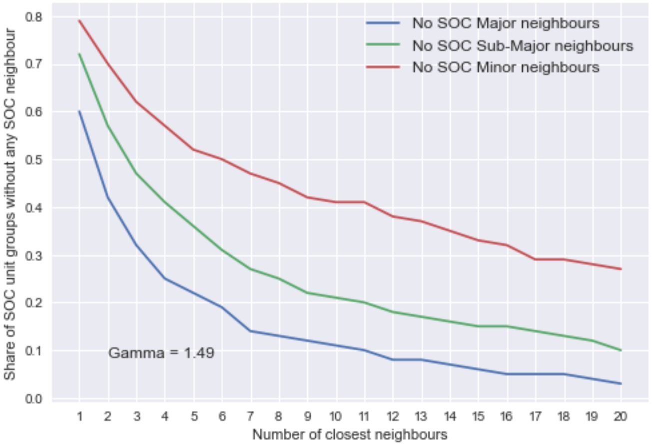
- RBF gaussian kernel $\kappa(x, y) = exp(-\gamma \mid\mid x - y \mid\mid^2)$
- small variance ($\gamma = \frac{1}{med^2} = 1.49$, with *med* the median of pairwise distances)

  γ obtained via the median heuristic (Muandet et al., 2017, 54)

Are the closest minor groups in the same (sub-) major group?



Are the closest unit groups in the same minor group?



Occupation classification

Occupations: **nine major groups** of the UK Standard Occupational Classification (SOC 2010)

1. Managers, directors and senior officials
2. Professional occupations
3. Associate professionals and technical occupations
4. Administrative and secretarial occupations
5. Skilled trades occupations
6. Caring, leisure and other service occupations
7. Sales and customer service occupations
8. Process, plant and machine operatives
9. Elementary occupations

Back