EXTENDED ABSTRACT: Enabling Workforce Intelligence through Occupational Taxonomy Alignment

Yifan Xu, Isabella Loaiza, César A. Uribe

Abstract

Labor market analysis and workforce planning increasingly rely on synthesizing data from multiple non-comparable occupational taxonomies. This fragmentation introduces substantial information loss, limiting the ability of governments, firms, and workers to measure equivalence and change in rapidly evolving labor markets, coordinate decisions in global organizations, and compare occupations across industries and local markets, particularly under conditions of technological change. We introduce a network-based occupational translator that aligns occupations across taxonomies by representing them as structured collections of tasks embedded in a global task meta-network. Task relationships are estimated using marginal co-occurrence or sparse inverse covariance estimation models, yielding alternative network representations that capture complementary notions of task interdependence. We formulate occupational dissimilarity as a Fused Gromov–Wasserstein problem that jointly aligns task semantics derived from textual embeddings and the structure of occupation-specific task networks. We apply this framework to align O*NET U.S. occupations from 2017 to 2024. The resulting translator recovers an approximately 80% agreement with official crosswalks. Overall, our preliminary results demonstrate how network structure provides a principled and interpretable foundation for occupational alignment beyond manual crosswalks.

1 Introduction

Occupational taxonomies are essential for labor market analysis, workforce planning, and economic measurement [1, 2]. Governments use them to monitor employment and productivity, firms to manage internal labor markets, and researchers to examine occupational mobility and structural change. However, these classification systems are periodically revised to reflect technological advancements, organizational changes, and shifting skill demands [3, 2]. While these updates are necessary, they complicate longitudinal analysis by disrupting direct comparability over time and across different contexts [3].

Crosswalks or manual mappings between occupational codes help with administrative continuity but are often coarse and opaque [4, 5]. They show equivalence mainly at the level of titles or categories, offering little insight into changes in actual job content [4]. They struggle to represent partial similarities, many-to-many relationships, or structural divergences in occupations [4]. Consequently, significant aspects of occupational change, like task reorganization or new task combinations, often remain unclear [6, 7].

An alternative perspective views occupations not as fixed labels but as structured collections of tasks and skills [6, 7, 8]. This approach represents occupations as networks, with tasks as nodes and their relationships as edges, highlighting patterns of interaction in labor markets [9]. *It also raises a key question: how should the similarity between two occupational networks be defined and measured?*

We develop a network-based framework for comparing occupations across taxonomies and over time by constructing a global task meta-network from textual task descriptions [10, 9]. This framework creates occupation-specific subnetworks to capture the organization of work, using both marginal co-occurrence networks and sparse inverse covariance methods to identify direct task dependencies [11, 12]. We employ an optimal transport-based distance that accounts for node attributes and network structure for comparing these subnetworks [13, 14, 15]. Applying this to align O*NET U.S. occupations from 2017 to 2024 [16], we achieve approximately 80% agreement with official crosswalks [4]. Our results illustrate how network structure provides a principled foundation for occupational alignment, enabling comparisons across diverse tasks while maintaining interpretability through soft alignments [15].

2 Methods

Data: We work with two datasets corresponding to distinct time periods: $\mathcal{O}_{2017} = \{\, o_1^{(2017)}, \dots, o_m^{(2017)} \,\}$, $\mathcal{O}_{2024} = \{\, o_1^{(2024)}, \dots, o_n^{(2024)} \,\}$ representing the sets of occupations in 2017 with $m = 959$ and 2024 with $n = 894$ [10], respectively. Each occupation o is associated with a set of textual task descriptions. Let $\mathcal{T} : \mathcal{O} \to \mathcal{P}(\mathcal{X})$ be a map from occupations to their respective subset of textual descriptions. Here, $\mathcal{X}$ is the complete set of task descriptions, with $|\mathcal{X}| \approx 17,000$, and $\mathcal{P}(\cdot)$ denotes the powerset. Note that each occupation can have multiple associated tasks, and hence multiple associated task descriptions.

We represent each task purely by its textual content using a pretrained sentence encoder. For a task description $x \in \mathcal{X}$, we compute a semantic embedding $E(x) \in \mathbb{R}^{768}$, where E denotes the SentenceTransformer model `all-mpnet-base-v2` [17]. The 768-dimensional vectors capture semantic similarity between task texts. Similarly, to obtain higher-level *task clusters*, we perform k-means clustering ($k = 750$) on the set of embeddings $\{E(x) : x \in \mathcal{X}\}$. Let $\mathcal{C} = \{c_1, \cdots, c_k\}$ be the set of cluster labels with the corresponding centroids $C = \{\mu_1, \cdots, \mu_k\}$. Let $\pi : \mathcal{O} \to \mathcal{P}(\mathcal{C})$ map each task to its assigned clusters. For each occupation o, with a slight abuse of notation, we define the corresponding set of task clusters as $\mathcal{C}(o) \triangleq \{\pi(x) \mid x \in \mathcal{T}(o)\} \subseteq \mathcal{C}$. Additional details on data set construction and preprocessing can be found in [9].

Task meta-networks. We model relationships between task clusters through a global task meta-network whose nodes are the cluster centers in C. We consider two complementary constructions: 1) We define a weighted undirected graph $G = (C, W)$, where the adjacency matrix $W \in \mathbb{R}_+^{k\times k}$ encodes marginal co-occurrence of task clusters across occupations: $W_{ij} = |\{\, o \in O : c_i \in C(o),\ c_j \in C(o) \,\}|$. This construction captures how frequently pairs of task clusters appear together in real-world job descriptions. 2) To isolate direct associations between task clusters, we construct a meta-network using sparse inverse covariance estimation (graph lasso). Let $X \in \mathbb{R}^{|O|\times k}$ be the occupation–cluster incidence matrix defined by $X_{o,i} \triangleq \mathbf{1}\{c_i \in C(o)\}$, and let S denote the empirical covariance matrix of the centered columns of X. We estimate a sparse precision matrix by solving the graphical lasso problem

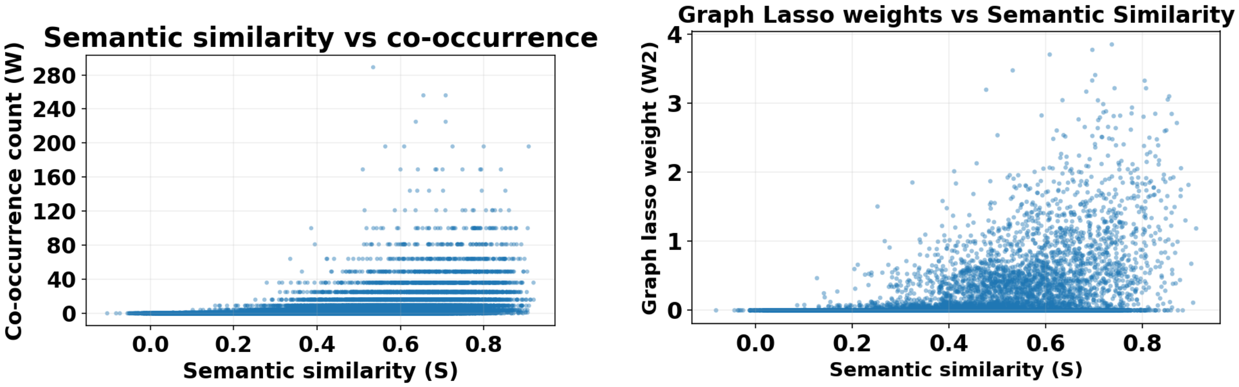
$$\widehat{\Theta} = \arg\min_{\Theta \succ 0} \langle S, \Theta \rangle - \log\det(\Theta) + \lambda \|\Theta\|_1,$$

where $\lambda > 0$ controls sparsity. Nonzero off-diagonal entries $\widehat{\Theta}_{ij}$ indicate conditional dependence between task clusters c_i and c_j after accounting for all others. We define the corresponding meta-network by edge weights $W_{ij}^{(2)} \triangleq |\widehat{\Theta}_{ij}|,\ i \neq j$, with $W_{ii}^{(2)} \triangleq 0$ for all i. Given either construction, each occupation o induces a task subgraph $G(o) = G[C(o)]$, obtained by restricting the global meta-network to the task clusters present in that occupation. These occupation-specific task networks encode not only which tasks are performed, but also how tasks relate to one another within an occupation.

Occupational Translator via Fused Gromov–Wasserstein distance. To compare occupations across years, we measure dissimilarity between their induced task networks using the Fused Gromov–Wasserstein (FGW) distance, which jointly aligns node features and network structure. In our construction, node features encode the semantic dissimilarity among the various task clusters. To do so, we construct the feature ground cost between clusters $M \in \mathbb{R}_+^{750\times 750}$ as the ℓ^2 norm between cluster centroids $M_{ij} \triangleq \|\mu_i - \mu_j\|_2 \ \ \forall i, j$.

For each occupation-specific task network $G(o)$, we define a structural cost matrix $C(o)$ that measures dissimilarity between pairs of nodes within the network based on their local connectivity patterns within the network. This captures how task clusters play comparable structural roles inside an occupation, independent of their semantic similarity. Now, we are ready to state our Fused Gromov–Wasserstein formulation. Let $o \in O_{2017}$ and $o' \in O_{2024}$, with node sets of sizes n and m, respectively. We place uniform probability mass on nodes, $p = \frac{1}{n}\mathbf{1}_n,\ q = \frac{1}{m}\mathbf{1}_m$, and let $\Pi(p, q)$ denote the set of feasible couplings between them. The FGW distance between o and o' is defined as

$$d_{\mathrm{FGW}}(o, o') = \min_{T \in \Pi(p,q)} \ \alpha \langle M(o, o'), T \rangle + (1-\alpha) \sum_{a,a',b,b'} \left| C_{aa'}^{(2017)}(o) - C_{bb'}^{(2024)}(o') \right|^2 T_{ab} T_{a'b'}, \qquad (1)$$



(a) Co-occurrence weights W with semantic similarity S. (b) Graph lasso weights $W^{(2)}$ with semantic similarity S.

Figure 1: **Task semantic similarity vs. task co-occurrence.** Scatter of cluster-pair W_{ij} for both methods versus cosine similarity S_{ij}.

where $\alpha \in [0, 1]$ balances semantic and structural contributions. In our experiments, we use $\alpha = 0.5$ as a balanced default and compute FGW distances using existing optimal transport solvers.

Given the FGW dissimilarity $d_{\mathrm{FGW}}(\cdot, \cdot)$ (1), our primary output is the cross-year dissimilarity matrix

$$D \in \mathbb{R}_+^{m \times n}, \qquad D_{ij} := d_{\mathrm{FGW}}\big(o_i^{(2017)},\, o_j^{(2024)}\big). \tag{2}$$

Matrix D defines a bipartite dissimilarity geometry between occupations across years, without imposing a one-to-one correspondence.

We view an *occupation translator* as an operator $\mathcal{T}$ acting on D to produce task-relevant crosswalk outputs at a chosen resolution: $\mathcal{T} : \mathbb{R}_+^{m \times n} \to \mathcal{Y}$, $\mathcal{T}(D) \in \mathcal{Y}$, where $\mathcal{Y}$ may represent, for example, ranked candidate matches for each source occupation (top-k retrieval), a soft alignment kernel obtained by row-wise normalization of $\exp(-\beta D)$, or a thresholded bipartite relation identifying all pairs with dissimilarity below a given level. Beyond pointwise matches, the optimal coupling $T^\star$ produced by (1) yields an alignment between task clusters for each pair of occupations, enabling task-level interpretation of occupational similarity and change.

3 Preliminary Results

Task semantics vs. task relatedness at the cluster level. We first validate whether the semantic geometry learned from task text aligns with the empirical organization of task clusters induced by our two task meta-network constructions. We represent each cluster by its centroid $\{\mu_i\}_{i=1}^k \subset \mathbb{R}^{768}$, and define semantic similarity via cosine similarity

$$S_{ij} = \frac{\mu_i^\top \mu_j}{\|\mu_i\|_2 \|\mu_j\|_2}, \qquad i, j \in [k].$$

In parallel, we measure task-relatedness either by (i) marginal co-occurrence weights W_{ij}, or (ii) conditional-dependence (graphical lasso) weights $W_{ij}^{(2)}$. We quantify the association between semantic similarity and meta-network connectivity using the Mantel test [18], computed after zeroing diagonals. Figure 1 shows that both constructions exhibit an overall increasing trend with higher semantic similarity. The co-occurrence weights W achieve a highly statistically significant positive correlation with S (Mantel Spearman $r = 0.4342$, one-sided $p = 0.001$). Moreover, despite the sparsity of the graph-lasso meta-network (6.60% nonzero off-diagonal entries versus 99.87% for S), we still observe a statistically significant positive association between $W^{(2)}$ and S (Mantel Spearman $r = 0.2932$, one-sided $p = 0.001$). These results indicate that both empirical occupation associations and estimated *direct* task dependencies are broadly consistent with the semantic geometry.

Although the positive association supports the view that the task meta-network reflects meaningful task-relatedness grounded in both language and observed co-appearance patterns, high-W_{ij} but moderate-S_{ij} outliers suggest the existence of cross-functional complements (tasks bundled together despite different semantic "themes"), whereas high-S_{ij} but low-W_{ij} pairs suggest near-substitutes or tasks that are semantically similar but rarely required jointly. These deviations motivate why a purely semantic alignment (e.g., nearest-centroid matching) can be insufficient: real occupations often combine complementary task clusters that are not maximally similar in embedding space, and capturing this requires information from the task-level subnetwork.

Cross-year occupational dissimilarity. We next examine the global geometry of cross-year occupational dissimilarity induced by FGW under the two task meta-network constructions. For each occupation pair $(o_i^{(2017)}, o_j^{(2024)})$, we compute the FGW distance $d_{\mathrm{FGW}}(o_i^{(2017)}, o_j^{(2024)})$ and assemble the full cross-year dissimilarity matrix D as defined in (2).

Figure 2 visualizes these matrices as heatmaps. Overall, the two constructions produce largely consistent FGW distances downstream. In particular, both heatmaps reveal a prominent band of high dissimilarity, corresponding to O*NET major group 25 ("Education, Training, and Library Occupations") in both years. We hypothesize that this effect is driven by differences in subnetwork size: occupations in group 25 are associated with substantially fewer tasks than the remainder of the taxonomy (mean 9.1 vs. 19.5), which yields smaller induced subnetworks and, consequently, larger FGW distances.

Both heatmaps also exhibit a clear block structure: FGW distances tend to be more homogeneous within major occupation groups and more variable across groups. Dark blocks along the main diagonal indicate low self-dissimilarity (i.e., occupations that remain close to their within-group counterparts across years) under both meta-network constructions.

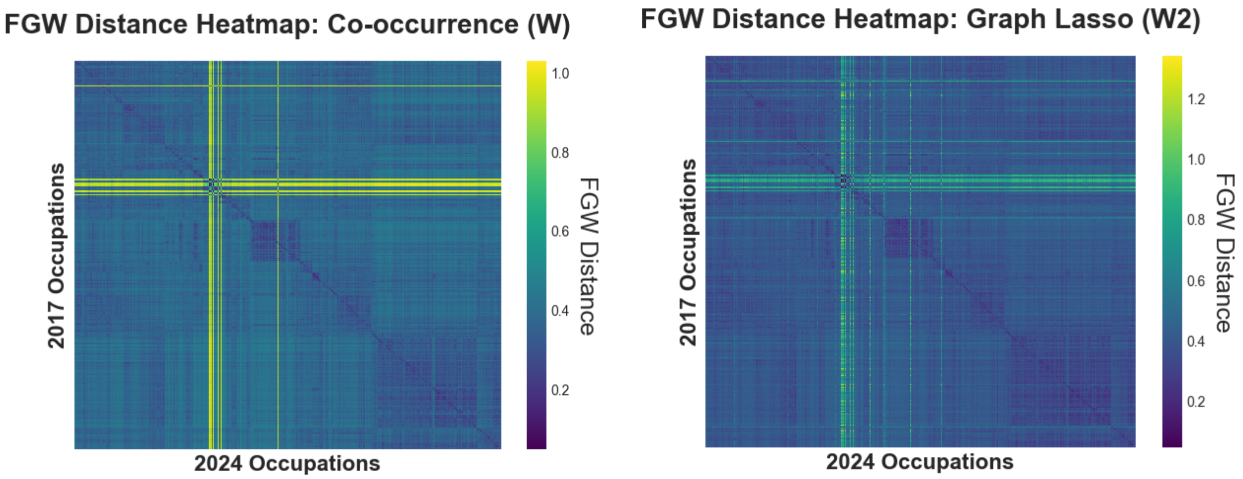
Despite this broad agreement, the two constructions differ in how they distribute contrast across the matrix. The co-occurrence-based heatmap yields more sharply separated dissimilarities, visible as brighter bands. In contrast, the graph-lasso-based heatmap highlights more localized irregularities, manifested as numerous thin high-dissimilarity stripes that suggest isolated occupation pairs behaving anomalously relative to their surrounding groups.

Taken together, these patterns indicate that the choice of task meta-network has a limited impact on the coarse geometry of D, but can meaningfully affect fine-grained structure. In practice, the appropriate construction should be selected to match the downstream occupation-translator operator (e.g., whether one prioritizes global separation for retrieval/thresholding or sensitivity to localized deviations for diagnostic analysis).

Translator reliability across occupational groups. Lastly, we evaluate our dissimilarity matrix D against existing data: a manually constructed occupation crosswalk between $\mathcal{O}^{(2017)}$ and $\mathcal{O}^{(2024)}$ [19]. Let $\mathfrak{c} : \mathcal{O}^{(2017)} \to \mathcal{P}(\mathcal{O}^{(2024)})$ denote this mapping.

Given a dissimilarity matrix D, for each $o^{(2017)} \in \mathcal{O}^{(2017)}$, let us denote $(o_1^{(2024)}, o_2^{(2024)}, \cdots, o_{|\mathcal{O}^{(2024)}|}^{(2024)})$ as the sequence of $\mathcal{O}^{(2024)}$ in non-decreasing order of their dissimilarity to $o^{(2017)}$. Then, we define the k-score of $o^{(2017)}$ as the minimum positive integer k such that $\mathfrak{c}(o^{(2017)}) \subseteq \{o_1^{(2024)}, \cdots, o_k^{(2024)}\}$. Intuitively, a smaller k-score indicates good agreement between the dissimilarity and the manual crosswalk, and vice versa. In particular, if an occupation $o^{(2017)}$ has a k-score of 1, then the $\mathfrak{c}(o^{(2017)})$ coincides with the most similar occupation under the FGW distance.

Overall, our dissimilarity achieves a k-score of 1 for 776 out of 959 occupations, with 894 attaining a k-score of at most 5. Figure 3 depicts the k-score distribution of the dissimilarity derived from graph lasso meta networks, organized by O*NET occupation group. Consistent with our findings in Figure 2, O*NET group 25 ("Education, Training, and Library Occupations") exhibits a large concentration of high k-scores and a median k-score of 10, indicating systematic disagreement between our dissimilarities and the manual crosswalk. Moreover, group 47 ("Construction and Extraction") has the highest median k-score, indicating that this group departs most strongly from the manual crosswalk. Understanding the sources of these within-group discrepancies is the subject of ongoing work.



(a) D built from W (co-occurrence).

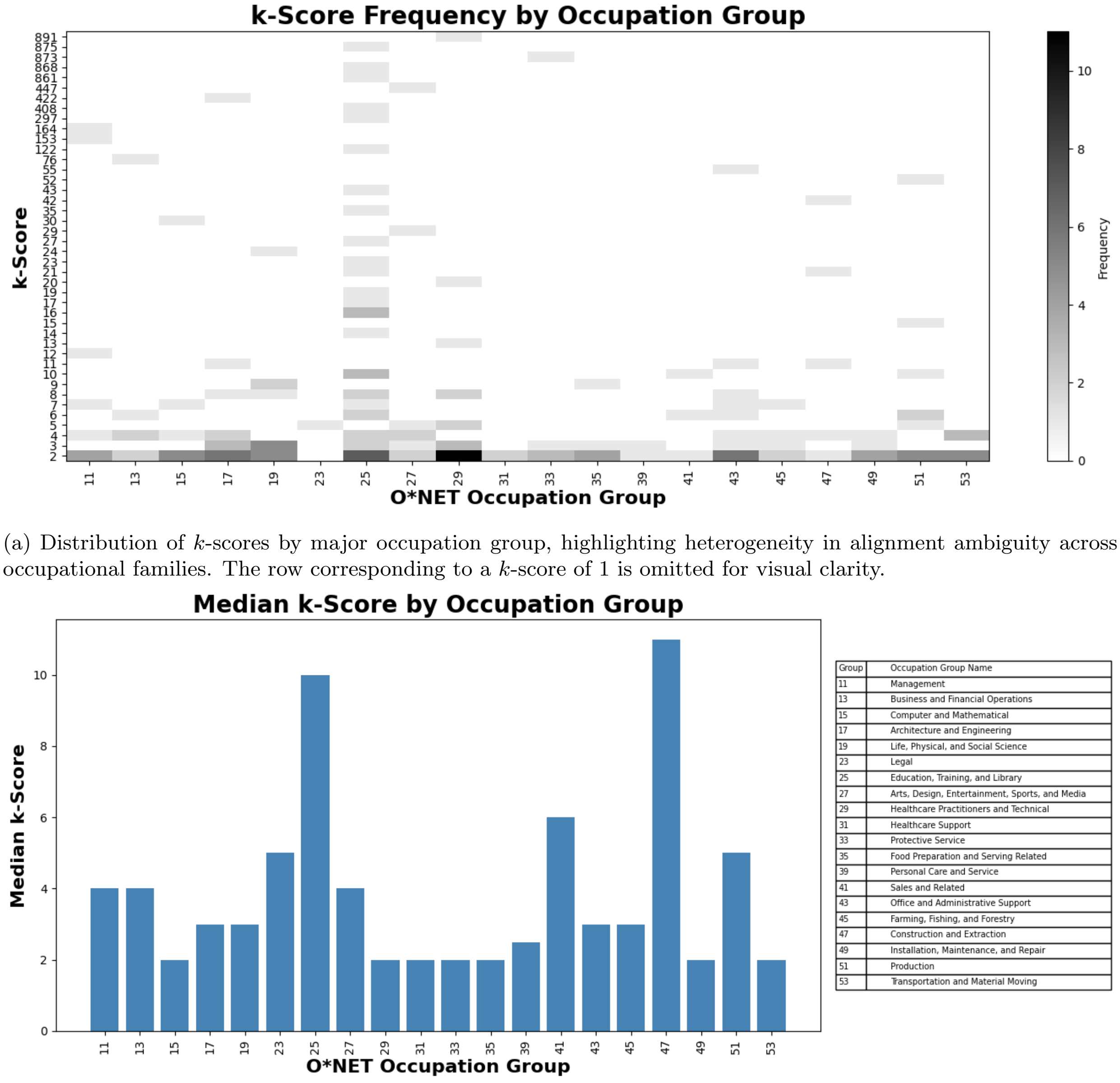
(b) D built from $W^{(2)}$ (graphical lasso).

Figure 2: **Cross-year occupational dissimilarity geometry.** Heatmaps of the cross-year FGW dissimilarity matrix D under two task meta-network constructions (co-occurrence W vs. conditional dependence $W^{(2)}$). Occupations are sorted based on their O*NET-SOC code in increasing order.

4 Conclusion and Future Work

We presented a network-based framework for aligning occupations across heterogeneous taxonomies by representing them as structured collections of tasks embedded in a global task meta-network. By integrating semantic information from task text embeddings with relational information captured through task co-occurrence and conditional-dependence networks, we reframed occupational comparison as a problem of network alignment rather than static code matching. Empirically, we showed that semantic similarity between task clusters is positively associated with their co-occurrence across occupations, and that Fused Gromov–Wasserstein alignment recovers substantial agreement with official crosswalks while revealing systematic deviations driven by differences in task organization.

These findings suggest several directions for future work. On the modeling side, richer task networks—such as weighted, directed, or temporally evolving graphs—could better capture asymmetric or dynamic task relationships. Methodologically, alternative network comparison tools or regularization strategies may further clarify the role of task interdependence in occupational similarity. From an applied perspective, extending this framework to cross-country taxonomies, more finely grained skill datasets, or worker-level transitions could support more detailed analyses of labor market change and workforce planning. More broadly, this work highlights the potential of network-based representations to support interpretable and flexible economic measurement in settings where traditional categorical mappings fall short.



(a) Distribution of k-scores by major occupation group, highlighting heterogeneity in alignment ambiguity across occupational families. The row corresponding to a k-score of 1 is omitted for visual clarity.

Group	Occupation Group Name
11	Management
13	Business and Financial Operations
15	Computer and Mathematical
17	Architecture and Engineering
19	Life, Physical, and Social Science
23	Legal
25	Education, Training, and Library
27	Arts, Design, Entertainment, Sports, and Media
29	Healthcare Practitioners and Technical
31	Healthcare Support
33	Protective Service
35	Food Preparation and Serving Related
39	Personal Care and Service
41	Sales and Related
43	Office and Administrative Support
45	Farming, Fishing, and Forestry
47	Construction and Extraction
49	Installation, Maintenance, and Repair
51	Production
53	Transportation and Material Moving

(b) Median k-scores for each occupational group, and a code-name mapping of the occupational groups.

Figure 3: FGW translator evaluation by occupational group

References

[1] *2018 Standard Occupational Classification Manual*. Executive Office of the President, Office of Management and Budget. 2018. URL: https://www.bls.gov/soc/2018/soc_2018_manual.pdf.

[2] International Labour Office. *International Standard Classification of Occupations: ISCO-08*. Geneva: International Labour Office, 2012. URL: https://webapps.ilo.org/ilostat-files/ISCO/newdocs-08-2021/ISCO-08/ISCO-08%20EN%20Vol%201.pdf.

[3] *2018 SOC User Guide*. Executive Office of the President, Office of Management and Budget. 2018. URL: https://www.bls.gov/soc/2018/soc_2018_user_guide.pdf.

[4] *2010 SOC to 2018 SOC Crosswalk: Explanatory Notes*. U.S. Bureau of Labor Statistics, 2018. URL: https://www.bls.gov/soc/2018/soc_note_2010_to_2018_crosswalk.pdf.

[5] Peter B. Meyer and Anastasiya M. Osborne. *Proposed Category System for 1960–2000 Census Occupations*. Tech. rep. U.S. Bureau of Labor Statistics, 2005. URL: https://www.bls.gov/osmr/research-papers/2005/pdf/ec050090.pdf.

[6] David H. Autor, Frank Levy, and Richard J. Murnane. "The Skill Content of Recent Technological Change: An Empirical Exploration". In: *The Quarterly Journal of Economics* 118.4 (2003), pp. 1279–1333. DOI: 10.1162/003355303322552801.

[7] Daron Acemoglu and David Autor. "Skills, Tasks and Technologies: Implications for Employment and Earnings". In: *Handbook of Labor Economics*. Ed. by Orley Ashenfelter and David Card. Vol. 4B. Elsevier, 2011, pp. 1043–1171. DOI: 10.1016/S0169-7218(11)02410-5.

[8] *The O*NET Content Model*. O*NET Resource Center / National Center for O*NET Development, 2025. URL: https://www.onetcenter.org/content.html.

[9] Humberto Loaiza and Roberto Rigobon. *The EPOCH of AI: Human-Machine Complementarities at Work*. Tech. rep. Last revised 2025-10-01. SSRN, 2024. DOI: 10.2139/ssrn.5028371.

[10] *The O*NET Database*. O*NET Resource Center / National Center for O*NET Development, 2025. URL: https://www.onetcenter.org/database.html.

[11] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3 (2006), pp. 1436–1462. DOI: 10.1214/009053606000000281.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441. DOI: 10.1093/biostatistics/kxm045.

[13] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *arXiv preprint* (2018). arXiv: 1803.00567.

[14] Facundo Mémoli. "Gromov–Wasserstein Distances and the Metric Approach to Object Matching". In: *Foundations of Computational Mathematics* 11 (2011), pp. 417–487. DOI: 10.1007/s10208-011-9093-5.

[15] Titouan Vayer et al. "Fused Gromov-Wasserstein Distance for Structured Objects". In: *Algorithms* 13.9 (2020), p. 212. DOI: 10.3390/a13090212.

[16] National Center for O*NET Development. *O*NET OnLine*. Web site. Accessed: January 26, 2026. 2026. URL: https://www.onetonline.org/.

[17] Kaitao Song et al. "Mpnet: Masked and permuted pre-training for language understanding". In: *Advances in neural information processing systems* 33 (2020), pp. 16857–16867.

[18] Nathan Mantel. "The detection of disease clustering and a generalized regression approach". In: *Cancer research* 27.2_Part_1 (1967), pp. 209–220.

[19] O*NET Resource Center. *Crosswalk 2010 to 2019 (Occupational Listings)*. Web page. Accessed 2026-01-30. 2019. URL: https://www.onetcenter.org/taxonomy/2019/walk.html.