UNIVERSITY OF COPENHAGEN

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Occupations as Structural and Symbolic Units

Computational Approaches to Labor Markets, Culture, and Assortative Marriage

Ella Clement

Supervised by Niels Richard Hansen, Andreas Bjerre-Nielsen and Sune Lehmann

2025-12-05

Ella Clement

Occupations as Structural and Symbolic Units

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen, 2025-12-05

Supervisors: Niels Richard Hansen, Andreas Bjerre-Nielsen and Sune Lehmann

University of Copenhagen

Faculty of Science

Department of Mathematical Sciences

Universitetsparken 5

2100 Copenhagen Copenhagen

Acknowledgements

Thank you to my supervisors, Niels Richard Hansen, Andreas Bjerre-Nielsen, and Sune Lehmann, for providing feedback and guidance throughout the course of my PhD.

Thank you to Michele Coscia and Lasse Mohr Mikkelsen for providing additional feedback and assistance on initial stages of specific projects.

Last but not least, thank you to Statistics Denmark for producing high quality datasets and making these available for research use. And in particular, thank you to Linnea Ranja Mignon Lanzky, Jolien Cremers, Laust Hvas Mortensen, Pernille Stender, and Emilie Rune Hegelund for help with my data access set-up and generously answering my questions about the data and Statistics Denmark infrastructure.

Abstract

Occupations are among the most fundamental units of social organization. They structure access to resources, shape identities, and serve as a primary medium through which inequality and meaning are reproduced. This dissertation examines occupations as both *structural positions* in the labor market and *symbolic resources* in culture using computational methods. It uses large-scale registry and cultural datasets to analyze how occupations are organized, portrayed, and paired across multiple empirical domains.

The first project develops a data-driven coarse-graining of the Danish labor market using monthly registry data from 2011–2022. By training a word2vec model on full-population career trajectories, the study embeds occupations in a continuous space defined by empirical mobility flows and identifies 311 occupation communities through density-based clustering. The results reveal a nuanced structure of labor market connectivity that extends beyond conventional classifications, capturing both expected professional domains and emergent hybrid clusters. These communities provide new insights into labor market dispersion and resilience to external shocks such as COVID-19.

The second project turns to cultural representations of work, assembling a dataset of 1,181 bestselling novels in the U.S. from 2023–2024. Using large language model (LLM)–assisted text analysis validated through manual review, it extracts the occupations of protagonists and romantic partners across genres. The findings show sharp divergences between fiction and reality: a narrow set of archetypal roles dominates, while many common occupations, especially in service and care work, remain largely invisible. Romantic portrayals are gendered, with men overrepresented in roles of power and risk and women

in expressive or domestic occupations. These representations illuminate how cultural products sustain selective visions of labor, value, and desirability.

The third, exploratory project integrates these two strands by examining assortative marriage through the lens of occupational distance. Using Danish registry data, it links married couples to their employment histories and measures occupational similarity via embedding-based distance metrics. Compared with traditional same-occupation indicators, this approach reveals graded patterns of proximity and assortative matching that extend across related occupations.

Methodologically, the dissertation demonstrates how computational techniques can be used as theory-generating tools in sociology. Embeddings, clustering, and LLM-assisted coding each extend the scale and granularity of analysis while maintaining interpretive rigor through careful validation and transparency. Substantively, the findings advance a unified framework for understanding occupations as both material and symbolic coordinates of social life: they organize economic opportunity and, at the same time, serve as cultural markers of identity and aspiration.

Taken together, the three projects show that understanding the modern world of work requires attention to both its structures and its stories. Computational methods, when critically employed, make it possible to analyze these dual dimensions in tandem, revealing how social hierarchies are built, represented, and reproduced across the intertwined realms of labor and culture.

Resumé

Erhverv er blandt de mest grundlæggende byggesten i den sociale organisering. De strukturerer adgangen til ressourcer, former identiteter og fungerer som et centralt medium, hvor ulighed og kulturel mening reproduceres. Denne afhandling undersøger erhverv som både *strukturelle positioner* på arbejdsmarkedet og *symbolske ressourcer* i kulturen ved hjælp af computational sociologiske metoder. Med store register- og kulturdatasæt analyserer den, hvordan erhverv organiseres, fremstilles og forbindes på tværs af flere empiriske domæner.

Det første projekt udvikler en datadrevet *coarse-graining* af det danske arbejdsmarked baseret på månedlige registerdata fra 2011–2022. Ved at træne en `word2vec`-model på fuldbefolknings karriereforløb skabes *embeddings* af erhverv i et kontinuert rum defineret af faktiske mobilitetsstrømme. Gennem tæthedsbaseret *clustering* identificeres 311 erhvervsklynger. Resultaterne afdækker en fintmærkende struktur af arbejdsmarkedets forbindelser, der rækker ud over de officielle klassifikationer og indfanger både etablerede faglige domæner og nye hybride grupperinger. Disse klynger giver et mere nuanceret billede af arbejdsmarkedets segmentering og dets modstandsdygtighed over for chok som COVID-19.

Det andet projekt undersøger kulturelle fremstillinger af arbejde ved at opbygge et datasæt af 1.181 bestsellerromaner i USA (2023–2024). Med LLM-assisteret tekstanalyse, efterfulgt af manuel validering, udtrækkes hovedpersoners og romantiske partneres erhverv på tværs af genrer. Fundene viser markante forskelle mellem fiktion og virkelighed: et snævert sæt arketypiske jobtyper dominerer, mens mange almindelige erhverv—særligt inden for service- og omsorgsarbejde—stort set er fraværende. De romantiske fremstillinger er tydeligt kønnede, med en overvægt af mænd i magtfulde eller risikofyldte roller og kvinder i udøvende, relationelle eller hjemlige erhverv. Disse mønstre belyser, hvordan kulturelle produkter opretholder selektive forestillinger om arbejde, værdi og attraktivitet.

Det tredje, eksplorative projekt forbinder de to foregående spor ved at undersøge assortativ ægteskabsdannelse gennem et mål for erhvervsmæssig afstand. Med danske registerdata kobles ægtefæller til deres beskæftigelsesforløb, og erhvervslighed måles med *embedding*-baserede afstandsmål. I forhold til traditionelle indikatorer afslører denne tilgang graderede mønstre af nærhed og assortativ matching, som strækker sig på tværs af beslægtede erhverv og dermed giver et mere kontinuert perspektiv på erhvervsmæssig sortering.

Metodisk viser afhandlingen, hvordan computational teknikker kan fungere som teorigenererende redskaber i sociologien. Embeddings, clustering og LLM-assisteret kodning udvider analysens skala og opløsning uden at gå på kompromis med fortolkningsdybde, takket være systematisk validering og gennemsigtighed. Substantielt bidrager afhandlingen til en samlet ramme for at forstå erhverv som både materielle og symbolske koordinater i det sociale liv:

de organiserer økonomiske muligheder og fungerer samtidig som kulturelle markører for identitet og aspiration.

Samlet viser de tre projekter, at forståelsen af det moderne arbejdsliv kræver opmærksomhed på både dets strukturer og dets fortællinger. Kritisk anvendte computational metoder gør det muligt at analysere disse to dimensioner i sammenhæng og dermed afdække, hvordan sociale hierarkier skabes, fremstilles og reproduceres i de sammenvævede sfærer af arbejde og kultur.

Contents

1 Introduction

1.1 Motivation

1.1.1 Societal Importance

Occupations are among the most enduring social categories through which modern life is organized. They define the structure of the labor market and anchor individuals' sense of social position, community, and self-worth. From a sociological perspective, occupations mediate the relationship between education, skills, and rewards; from an economic one, they shape wage distributions, labor allocation, and the diffusion of technological change (Erikson and Goldthorpe, 1993; Acemoglu and Autor, 2011). In cultural terms, they also provide much of the symbolic vocabulary through which people narrate their lives (Dunkerley, 1975).

Occupational mobility—how individuals transition from one occupation to another—reveals the broader logic of opportunity in a society. Patterns of movement across occupational boundaries signal the permeability of class structures and the adaptability of economies to change (Blau and Duncan, 1967). Understanding these patterns is essential for assessing economic resilience: whether workers can realistically retrain, or whether labor markets reproduce inherited hierarchies (Kalleberg, 2009; Cahuc *et al.*, 2014).

These questions are particularly salient today, when workers move through jobs faster than ever before (Bidwell, 2013). Technological automation continues to reshape demand for skills, creating both new occupations and structural redundancies (Mäkelä and Stephany, 2025). The COVID-19 pandemic further exposed the fragility of occupational structures, amplifying inequalities between sectors that could transition to remote work and those that could not (Nwosu *et al.*, 2022). Together, these developments underscore that under-

standing occupational mobility is not merely an academic concern: it speaks directly to the capacity of societies to grow and adapt to change.

This thesis approaches occupations as a central axis of modern social structure, examining how they can be quantified, grouped, and compared across economic, cultural, and interpersonal dimensions. By combining large-scale administrative and cultural representation data with computational methods, it contributes to understanding how occupations both reflect and reproduce broader social patterns.

1.1.2 Scientific Opportunity

The study of occupations has long been central to sociology and labor economics, yet its empirical foundations have been limited by how occupations can be measured and compared. Traditional surveys and occupational classifications provide essential structure but offer only coarse, static snapshots of a dynamic social reality. Today, two major developments—the rise of population-scale administrative registries and the maturation of computational text analysis—open new frontiers for studying work.

Danish administrative registries exemplify the first development. Their population-wide, longitudinal coverage enables the reconstruction of complete employment histories, allowing occupations to be observed as evolving trajectories rather than single states. This makes it possible to examine social mobility, labor-market clustering, and assortative matching with unprecedented granularity.

The second development lies in the cultural domain. Advances in large-scale text processing and natural language modeling now allow sociologists to extract systematic evidence from narrative and media sources once accessible only through close reading. This offers a new way of studying cultural products at scale.

Together, these data and methodological innovations bridge two traditions: the structural analysis of labor markets and the cultural analysis of meaning-making. Embedding models and network methods translate occupational relationships into continuous spaces of similarity and mobility, while LLM-

assisted text analysis extends this representational logic to cultural narratives. The convergence of these tools creates a genuine scientific opportunity to revisit classical social scientific questions about inequality, status, and identity with data and methods that better capture contemporary work.

1.1.3 Central Tension and Gap

Historically, empirical research into occupational mobility has relied on fixed taxonomies, survey data, or prestige scores (Blau and Duncan, 1967; Ganzeboom, 2010). These frameworks have yielded foundational insights into class reproduction and labor market segmentation, but they depend on pre-defined coarse occupational categories and relatively small samples.

More recently, data-driven methods have come to the fore. These leverage computational modeling and large-scale behavioral data to infer relationships between occupations from observed transitions. These approaches reveal how individuals actually move, rather than how occupations are assumed to relate.

While data-driven modeling is an increasingly common way of studying occupational mobility, most existing studies consider only the most recent job transition and omit key labor market dynamics such as self-employment, part-time work, and longer-term career histories—segments which are essential for understanding labor precarity, self-organization, and the diversity of modern career paths. Moreover, many empirical studies continue to rely on smaller-scale surveys, which are potentially biased due to imperfect sampling, poor coverage, and missingness.

A related body of work examines the prestige of occupations, finding that cultural ideals of prestige in work are remarkably stable and cross-culturally valid (Treiman, 1977). Yet studies in this vein have also tended to rely on survey data, which are additionally prone to social desirability bias (Bound *et al.*, 2001). They also fail to disentangle the multi-faceted nature of prestige—what is prestigious for one's self is not necessarily prestigious in a romantic interest, for example. Additionally, analyses of occupational prestige and desirability do not extend to how such roles circulate through cultural products. Nor have these cultural portrayals been systematically compared to real labor

market data, leaving open questions about how media reflect, distort, or reinforce occupational hierarchies.

This dissertation positions itself at the intersection of these traditions. By combining structural labor market analysis with computational modeling and large-scale text analysis, it bridges economic and cultural perspectives on work. The goal is not only to map how occupations relate empirically within registries and career data, but also to understand how they are imagined and represented within the cultural sphere.

1.2 Overarching Research Question

Occupations function as both *structural* and *symbolic* units of social life. As structural units, they locate individuals within the organization of the labor market, shaping access to resources, mobility pathways, and patterns of inequality. As symbolic units, they carry cultural meanings that signal identity, value, and social distinction. Social scientific research has long examined these dimensions, but typically in separate traditions—one focused on occupations as positions in stratification and mobility structures, and another on how work is represented, narrated, and imbued with meaning. Far less is known about how these structural and symbolic dimensions relate to one another, or how they diverge across empirical domains.

At the same time, new computational methods make it possible to model occupations across both registers: as nodes in networks of mobility, as linguistic objects in text, and as vectors in continuous representational spaces. This creates an opportunity to integrate structural and cultural analyses of work in ways that were previously difficult to achieve.

This thesis thus addresses the question:

> **How can computational methods be used to advance social scientific understanding of occupations as both structural and symbolic units of social life?**

Below I describe how the three projects which make up this dissertation relate to this broader research question.

1.2.1 Research Question 1: Data-Driven Structure of Labor Markets

The first part of this thesis investigates how data-driven methods can uncover the latent structure and dynamics of labor markets beyond traditional classifications. Classical approaches to occupational analysis rely on expert-defined categories, which impose boundaries that may not reflect how people actually move through work. By contrast, computational models based on observed career transitions reveal the empirical structure of the labor market.

Manuscript 1 (Chapter 3) addresses this question by applying embedding and clustering methods to Danish registry data, representing occupations in a continuous vector space derived from observed job-to-job transitions. Using these embeddings, the analysis identifies cohesive occupational communities that capture both expected and previously obscured relationships between jobs. This approach allows for an empirically grounded classification that complements official taxonomies while capturing worker histories, and scales to the size and complexity of a full labor market more effectively than previous methods.

The resulting framework also provides the foundation for later analyses in the thesis. The embedding-based similarity measures developed here are used in the exploratory Project 3 (Chapter 5) to model occupational assortative marriage, demonstrating that the same empirical representation of labor market structure can illuminate both individual career mobility and patterns of social reproduction.

1.2.2 Research Question 2: Cultural Representations of Occupations

The second research question examines how occupations are represented symbolically in cultural products, and how these portrayals align with real-

world labor markets and social desirability. While labor market data reveal the structural organization of work, cultural narratives reveal how societies imagine and evaluate that work. Fictional portrayals of occupations function as a mirror—sometimes distorted—of economic hierarchies, signaling which forms of labor are seen as respectable, romantic, or aspirational.

Manuscript 2 (Chapter 4) addresses this question by using large language model–assisted coding to identify and classify occupations across contemporary bestselling novels. The study systematically quantifies which professions appear most frequently, how they differ by genre, and how they are distributed across gendered character roles in romantic portrayals. By comparing these cultural distributions with real-world labor market data, the analysis highlights discrepancies between representation and reality: certain forms of professional or creative work are vastly overrepresented, while routine, care, and manual occupations remain largely invisible.

1.2.3 Research Question 3: Refining Measures of Assortative Marriage

The third research question explores how computational approaches can refine classical measures of *homogamy*, i.e., the marriage of individuals who are similar to one another, in this case by profession. Traditional studies of occupational homogamy rely on binary indicators of same-occupation pairing or on one-dimensional status scores (Schwartz *et al.*, 2021). While these measures have provided foundational insights, they overlook the nuanced degrees of relatedness between different jobs and the structural pathways that connect them within the labor market.

Chapter 5 builds on the embedding framework developed in Chapter 3 to address this limitation. By representing each occupation as a point in a continuous vector space derived from real job-to-job transitions, it becomes possible to measure occupational similarity continuously rather than categorically. This allows for fine-grained analysis of how couples form across related occupations, rather than just identical ones.

This research question thus connects the methodological innovation of Chapter 3 with the sociological theme of Chapter 4.

1.3 Background and Context

1.3.1 Occupations in Sociology

Since the mid-twentieth century, scholars have treated occupations as compact proxies for the complex constellation of education, income, and prestige that defines social structure (Blau and Duncan, 1967). Because occupational titles capture both economic function and social standing, they provide a stable and comparable unit for examining inequality across time and space.

One strand of work focuses on occupational prestige. Prestige scales derived from public opinion surveys, such as those developed by Treiman (Treiman, 1977) and later updated by Ganzeboom (Ganzeboom and Treiman, 1996; Ganzeboom, 2010), established consistent hierarchies of occupational standing across diverse contexts. These scores made it possible to treat occupations as continuous indicators of social status, facilitating the study of intergenerational mobility, status attainment, and the reproduction of inequality (see e.g., Mazumder and Acosta, 2015; Ganzeboom *et al.*, 1991).

A second strand conceptualized occupations as categorical building blocks of *class schemas*. Frameworks such as the Erikson–Goldthorpe–Portocarero (EGP) schema and Oesch's class typology organized occupations into groups reflecting employment relations, skill specificity, and authority structures (Erikson *et al.*, 1979; Oesch, 2006). These schemas sought to move beyond purely prestige-based rankings by capturing differences in employment contracts and work logic—distinguishing, for example, between salaried professionals, routine employees, and manual workers. Within this view, occupations are embedded in a broader system of labor market positions that structure opportunities and constraints.

A crucial empirical link between these traditions was provided by the study of *mobility tables*. Pioneered in the mid-twentieth century, these cross-tabulations of origin and destination occupations formed the backbone of early research

on social mobility (Blau and Duncan, 1967). Mobility tables quantify how often individuals move between occupational categories relative to chance, thereby revealing the permeability of class boundaries and the persistence of inequality across generations. Yet these analyses required a theoretically meaningful set of categories—precisely what the EGP schema supplied. In this sense, EGP and later refinements were designed as interpretive frameworks for mobility tables, ensuring that measured mobility reflected underlying class structures rather than arbitrary occupational codes.

Oesch's typology updated the EGP approach for post-industrial labor markets, distinguishing technical, organizational, and interpersonal work orientations within the same broad employment relations. It thus captured the growing horizontal differentiation of contemporary class structures while maintaining the relational principles that made EGP analytically powerful.

More recent approaches extend this relational logic beyond discrete tables, treating occupations as nodes within networks of mobility or skill transfer (Cheng and Park, 2020). Network-based and embedding-based models can be seen as direct successors to the mobility-table tradition: where earlier work tabulated movement between aggregated class categories, these models infer structure from individual-level transitions across thousands of occupations, capturing the same principles at finer granularity.

1.3.2 Occupation Taxonomies

Sociological and policy-oriented analyses of work often rely on expert-defined classification systems such as the International Standard Classification of Occupations (ISCO) and the Statistical Classification of Economic Activities in the European Community (NACE), which respectively categorize occupations and industries (International Labour Office, 2023; Eurostat, 2008). These frameworks serve essential institutional functions: they enable standardized reporting of employment statistics, facilitate cross-national comparisons, and provide a shared language for describing the organization of work. Their design reflects decades of theoretical and bureaucratic effort to render the labor market legible as a coherent hierarchy of roles.

Within sociology, these descriptive systems exist alongside analytical class schemas such as the Erikson–Goldthorpe–Portocarero (EGP) classification and Oesch's typology of employment relations. While ISCO and NACE are organized by task and industry, EGP and Oesch explicitly model the social and economic relations embedded in employment—authority, autonomy, and market dynamics. Both traditions seek to impose order on the diversity of work, yet they rest on different logics: one functional and administrative, the other relational and sociological.

In practice, the two rarely align neatly. For instance, jobs with similar ISCO codes may fall into distinct EGP classes if they differ in contract type or managerial authority, while different ISCO codes may cluster together under Oesch's typology if they share similar employment relations. This divergence underscores how the chosen taxonomy shapes what aspects of the labor market become visible.

Despite their utility, these taxonomies have important limitations. They are inherently static: categories are revised infrequently, while the labor market itself evolves continuously. Emerging occupations—particularly those created by technological change, platform work, or hybrid skill sets—are often forced into legacy codes that inadequately describe their content or context. Expert-defined groupings also conceal heterogeneity within categories: roles with the same code may differ sharply in working conditions, prestige, or gender composition, while related occupations may be divided by arbitrary taxonomic boundaries. Moreover, the ISCO and NACE frameworks, though complementary, are difficult to align, covering non-overlapping domains with distinct coding rules.

They also, by design, do not express the cultural and symbolic dimensions of work. They classify by task and relation, not by meaning. Consequently, they cannot capture how occupations are valued, imagined, or narratively represented—dimensions that matter for understanding how work figures in identity and culture.

1.3.3 Computational Approaches

The last decade has seen a rapid expansion of computational approaches to studying work, drawing on large-scale surveys or digital traces and advances in natural language processing and network science. These methods build upon—but also transcend—the classical tools of occupational sociology. Where traditional analyses relied on pre-defined taxonomies or aggregated mobility tables, computational techniques allow for data-driven modeling of the latent structure of work empirically.

Network-based models have generalized the logic of the mobility table, representing the labor market as a graph of occupational transitions rather than a set of discrete categories. Nodes correspond to occupations, and edges represent empirically observed flows between them. Such representations make it possible to identify clusters or ”communities” of jobs that are functionally interconnected, even when they span formal classification boundaries (Sørensen and Grusky, 1996; Cheng and Park, 2020). These methods have been used, among other purposes, to analyze the permeability of class boundaries, assess automation risk, and reveal how skill adjacency shapes economic resilience (Rio-Chanona *et al.*, 2020; Escobari *et al.*, 2021; Villarreal, 2020).

In parallel, advances in natural language processing (NLP) have introduced powerful new tools for representing life events as continuous vectors in semantic space. Embedding models such as word2vec, fastText, and more recently transformer-based encoders, learn representations from large corpora, and have been applied to both text and tabular data. These embeddings can be trained to capture fine-grained relationships between life events or outcomes. In the context of occupations, studies have translated free-text job descriptions into standardized codes (Kim *et al.*, 2024; Dahl *et al.*, 2024), modeled career trajectories using sequential embeddings (Vafa *et al.*, 2024), and predicted life outcomes from registry data, including occupational data (Savcisens *et al.*, 2024).

Computational methods have also begun to enable large-scale analysis of cultural products. Recent advances in natural language processing and computer vision have made it possible to quantify themes, character types, and

social representations across massive corpora of novels, films, and other media. In cultural sociology, this has given rise to new strands of research that treat fiction, journalism, and entertainment not as anecdotal reflections of society but as analyzable data sources in their own right (Macanovic, 2022; Siddiqui, 2024). Large language models and embedding techniques can extract structured information from unstructured text, while topic models and semantic networks reveal how symbolic associations cluster and evolve over time (Chhetri *et al.*, 2025; Steyvers and Tenenbaum, 2001; Gerlach *et al.*, 2018). These methods have been used to study the diffusion of narratives across genres (e.g., Schöch, 2021) and the relationship between cultural and economic hierarchies (e.g., Kozlowski *et al.*, 2019).

Despite their promise, these methods also raise new challenges. Embedding spaces can be difficult to interpret; data-driven models risk reifying biases present in the data; and the use of large administrative and textual corpora introduces reproducibility and privacy concerns (Simhi and Markovitch, 2023; Ntoutsi *et al.*, 2020). Additionally, different modeling assumptions are more suitable to different data, both in terms of data collection and scaling, so care must be taken to select appropriate methods. Nevertheless, when carefully validated and theoretically grounded, computational approaches can complement traditional frameworks.

1.3.4 The Danish Context

The methodological advances outlined above gain particular traction in contexts where labor markets are well-documented and institutions are stable. Denmark exemplifies such a context, combining rich administrative data with distinctive institutional arrangements that shape both mobility and inequality.

The Danish labor market is structured via policy according to a ”flexicurity” model, which combines relatively easy hiring and firing with generous unemployment insurance. The result is a system with both high security for workers and high flexibility for firms. It is also characterised by two further features: a large public sector that employs a substantial share of the workforce, and a compressed wage distribution produced by coordinated collective bargaining (Andersen, 2023). Together, these institutional arrangements pro-

duce low unemployment rates by international standards—2.9% as of August 2025 (Zimmermann, 2025)—alongside short unemployment spells and high levels of job-to-job mobility.

Female labor force participation is also high: in 2023 about 74% of women aged 20–64 were employed, well above EU averages (EURES (EURopean Employment Services), 2025). This dual-earner norm, supported by public childcare and parental leave, means that career dynamics and household formation unfold under conditions quite different from countries where women participate less or where unemployment is persistent.

Denmark also keeps high-quality records on a large number of facets of residents' lives via the Civil Personal Register, established in 1968 to streamline taxation and administration. This system assigns every resident a unique personal identification number that links across all public systems (Pedersen, 2011). Because this ID appears in education, earnings, health, and employer registers, researchers can follow individuals, couples, and firms across time with complete coverage. The data enable population-scale, longitudinal, linkage-rich designs rarely possible elsewhere. In practice, this means it is feasible to reconstruct month-by-month occupational histories for millions of workers, or to trace the life-course trajectories of entire birth cohorts.

These institutional features shape both what can be learned from Danish data and what will or will not generalize. The flexicurity model, compressed wage distribution, and large public sector mean that occupation transitions are relatively frequent and less tied to wage shocks than in more rigid or unequal markets. This makes Danish data ideal for constructing fine-grained maps of occupational similarity: embeddings and clusters derived from transition graphs are dense and stable, reflecting skill and task proximity more than income gradients. In countries with lower mobility or higher inequality, however, the same methods may produce sparser or more wage-driven clusters. Similarly, assortative marriage by occupation in Denmark occurs in a setting where most couples are dual earners; in settings with lower female participation or larger informal sectors, part of what looks like assortative marriage may instead reflect selection into employment itself.

1.3.5 Cultural Representations of Work

Fiction is not only a source of entertainment but also a reflection and selective amplifier of cultural norms (Couldry, 2012). Occupations in novels, films, and television function as narrative conveniences precisely because they signal who characters are, what they value, and how they are positioned within social hierarchies. As scholars of work and culture have shown, such representations both reflect and reproduce the hierarchies of prestige, gender, and worth that structure real labor markets (Holtzman and Sharpe, 2014).

Cultural sociology and media studies have long examined how professions symbolize broader social values. Classic analyses link occupational archetypes to moral authority and modern identity: the doctor as rational healer, the teacher as nurturer of civic virtue, the police officer as agent of order (Larson, 1977). In genre fiction especially, occupations often function as shorthand for narrative roles—scientists in science fiction, soldiers in war stories, or detectives in mysteries (Cawelti, 1976). These archetypes not only express cultural ideals but also help naturalize particular divisions of labor as meaningful and desirable. Scholars of work have also emphasized the invisibility of many forms of everyday labor in cultural representation, especially service and care work (Glenn, 1992; England, 2005). The result is a systematic underrepresentation of the occupations that sustain daily life but lack glamour or narrative tension.

1.3.6 Homogamy

A long-standing theme in sociology is that people tend to form romantic partnerships and have children with others similar to themselves—a phenomenon known as *homogamy* or *assortative matching*[1]. This pattern has been studied across multiple dimensions, including education, income, and occupation (Kalmijn, 1998). Because partnerships often shape both household income and intergenerational transmission of advantage, assortative marriage is a central mechanism through which inequality is maintained or amplified over time (Luo, 2017).

[1] *Assortative mating* is another common term, but it is less accurate for my case, as I focus on marriage rather than childbearing.

Within this literature, occupations have served as a key indicator of social position. Occupational homogamy, where partners share the same or closely related jobs, has been used to infer social closure, shared values, and overlapping social networks (McPherson *et al.*, 2001; Kalmijn, 1998). Traditional measurement approaches fall into several categories. The simplest compares whether partners hold exactly the same occupation, treating similarity as a binary outcome (same versus different), and often focusing on a specific field (Schwartz *et al.*, 2021). Alternative approaches compare the status or prestige scores of partners' occupations, e.g., Kalmijn (1994).

While these approaches have yielded robust findings, they also face limitations. Binary indicators are too coarse to capture the meaningful similarities between distinct but related occupations—for example, between a nurse and a medical technician, or between a journalist and an editor. Conversely, they may overstate proximity between occupations culturally stereotyped as belonging together. Prestige- or class-based measures, meanwhile, collapse occupational differences into one-dimensional hierarchies, overlooking the relational and multidimensional nature of work. Few studies account for the actual mobility pathways between occupations—that is, whether transitions between two roles are common or rare, and thus whether they represent plausible meeting grounds.

1.4 Overview of Manuscripts

1.4.1 Manuscript 1: Data-driven Coarse-graining of Occupations

The first manuscript, presented in Chapter 3, investigates how data-driven methods can uncover latent occupational structures and labor market dynamics at a population scale. Whereas traditional classifications impose *a priori* ideas onto labor market structure, and most computational approaches have been developed on smaller or survey-based datasets, this study applies scalable embedding and clustering techniques to full-population registry data. The central question is how occupations cluster empirically, as revealed through patterns of observed career mobility, rather than formal definitions.

The analysis draws on the Danish Labour Market Account (Stender *et al.*, 2015), which began in 2008 and provides monthly employment records for the entire Danish population. Restricting my attention to the period 2011–2022 and selecting individuals aged 15–65 who held at least one job or self-employment spell during this period yielded a population of approximately 4.3 million workers. Each individual's sequence of occupations across months was treated as a career "sentence," allowing the use of word embedding techniques originally developed in natural language processing. A word2vec model (Mikolov *et al.*, 2013) was trained on these sequences, producing a continuous vector representation of each occupation such that jobs frequently held in succession are positioned closer together in the embedding space.

To identify broader patterns of occupational similarity, the embedding vectors were clustered using a projection-based density algorithm (HDBSCAN) (Thrun and Ultsch, 2021), yielding 311 empirically derived occupation communities. These communities varied systematically in their demographic, educational, and economic profiles. Some aligned with formal educational tracks—such as a "legal" community encompassing lawyers, judges, and legal assistants—while others captured more transient or life-stage-specific employment patterns, such as low-wage, short-duration jobs commonly held by workers under 25.

The manuscript demonstrates two use cases of the resulting occupational map. First, it traces the post-graduation diffusion of educational cohorts through the labor market, revealing how graduates from the same field disperse or remain clustered over time. Second, it identifies the communities most persistently affected by the COVID-19 pandemic and subsequent lockdowns, showing how churn in these communities remained high even after the end of lockdowns.

The contribution of this manuscript is both substantive and methodological. Substantively, it reveals how empirical occupational communities can illuminate structural inequalities and resilience within the labor market. Methodologically, it expands the sociological toolbox for large-scale, data-driven coarse-graining of occupations—offering a scalable alternative to expert-based classifications. The embedding framework developed here also forms the foundation for the exploratory study of homogamy in Chapter 5.

1.4.2 Manuscript 2: Occupations in Contemporary Popular Fiction

The second manuscript, presented in Chapter 4, examines how occupations are represented in contemporary fiction and how these portrayals align—or fail to align—with real-world labor markets and patterns of occupational desirability.

To address this question, I compiled a corpus of 1,181 bestselling adult novels published or widely circulated during 2023–2024 in the U.S. Using a large-language-model (LLM)–assisted pipeline, I extracted protagonist and (where relevant) love interest occupations and verified the results through manual validation of a 10% random subset. This hybrid approach combines computational scalability with interpretive reliability, allowing for systematic coding across a large and heterogeneous corpus.

The analysis compared the resulting distribution of fictional occupations with real-world occupational frequencies, prestige rankings, and data on partner desirability. The findings reveal patterned distortions in how work is represented. Certain archetypes—for instance students, writers, detectives, athletes—are strikingly overrepresented, while everyday occupations such as clerical, service, or care work are rarely depicted. High-prestige occupations are more often depicted, although they are not necessarily central: some, like engineers or tech executives, appear infrequently despite their perceived high social status, while others with lower real-world status persist as romantic or aspirational symbols. Gender patterns in romance are pronounced. Male characters are disproportionately associated with professions emphasizing power, danger, or authority (e.g., CEOs, mafia bosses), while female characters appear more often in roles tied to expressiveness, domesticity, or creativity (e.g., florists, artists). These asymmetries suggest that fiction mirrors but also selectively exaggerates cultural ideals of masculinity and femininity in work.

Methodologically, the manuscript demonstrates how LLM-assisted coding can make large-scale cultural analysis feasible while preserving qualitative nuance. Substantively, it contributes to the sociology of work by shifting attention from occupational prestige as attitude to occupational visibility as representation. In doing so, the study bridges labor-market sociology and cultural sociology,

providing systematic evidence of how fiction constructs the social imagination of work and desire in contemporary mass culture.

1.4.3 Exploratory Study: Assortative Marriage by Occupational Distance

The exploratory study presented in Chapter 5 extends the methodological framework developed in Chapter 3 to the domain of family and stratification research, focusing on assortative marriage. The central question is how one can use continuous occupational similarity measures to expand our view of assortative marriage by occupation beyond binary identities.

This study uses Danish registry data linking spouses and registered partners to their occupations, focusing on the same 2011–2022 period analyzed in Chapter 3. Each partner's occupation is represented by an embedding vector derived from the word2vec model trained on population-wide job transitions. Occupational similarity between partners is then quantified as the cosine similarity between their respective embeddings. To evaluate the distinctiveness of real-world patterns, observed couples are compared against a baseline of randomly recombined pairs holding the same marginal distributions of sex, age, and occupation frequency. This comparison isolates the extent to which partner matching reflects genuine occupational proximity rather than random pairing or compositional effects.

Preliminary results indicate that embedding-based distance measures yield a nuanced picture of occupational homogamy. Rather than a sharp divide between identical and different occupations, the data reveal a gradient of similarity: couples tend to pair within adjacent occupations.

The contribution of this exploratory project is twofold. Methodologically, it demonstrates the value of continuous embedding-based similarity metrics for refining classical sociological measures of homogamy. Substantively, it extends the labor market framework developed in Chapter 3 into the study of intimate relationships, illustrating how structural positions in the economy shape the social fabric of family formation.

1.5 Data Sources and General Approach

1.5.1 Registry Data (Denmark, 2011–2022)

Much of the empirical foundation for this thesis rests on Danish administrative registries, which provide an unparalleled resource for studying labor market structure and dynamics. Denmark's registry system covers the entire resident population, linking individuals across employment, income, education, and demographic domains through anonymized personal identifiers. This comprehensive and longitudinal structure enables population-level analyses that are both detailed and reliable.

The primary data source is the Danish Labour Market Account (LMA) (Stender *et al.*, 2015) compiled by Statistics Denmark, which I analyze in the years 2011–2022. The LMA provides monthly employment records encompassing roughly six million individuals, of whom about 3.8 million are of working age at a given time (Statistics Denmark, 2025). Each observation contains information on employment status, employer, occupation (coded according to a Danish tailoring of the ISCO-08 taxonomy), monthly hours worked in the occupation, among other information. Because the dataset is constructed through administrative reporting, it captures employment transitions with high accuracy and minimal recall bias.

These features make the LMA well-suited for analyzing occupational mobility and structure. Its monthly granularity allows researchers to observe not only job changes but also temporary exits from the labor market, self-employment, and part-time work—dimensions that are often missing or underreported in conventional data sources. The coverage of the entire Danish workforce enables analyses that scale to millions of career sequences, making it possible to model the empirical relationships among thousands of occupations in fine detail.

In addition to the LMA, supplementary Danish registries were used where relevant to specific analyses, including datasets on income, education, and demographic characteristics. Together, these registries form an integrated data infrastructure that supports longitudinal and population-level research

while ensuring data security and confidentiality through Statistics Denmark's controlled research environment. This combination of scale, precision, and continuity provides an exceptional foundation for the computational and sociological analyses undertaken in Manuscripts 1 and 3.

1.5.2 Fiction Dataset (2023–2024 Bestsellers)

To complement the structural labor market analyses, this thesis introduces a novel dataset capturing how occupations are represented in contemporary fiction. The dataset comprises 1,181 bestselling novels widely circulated between 2023 and 2024, spanning a range of popular genres. Titles were drawn from multiple adult fiction bestseller lists on Publishers Marketplace to ensure commercial success and broad readership across the U.S. market.

The corpus was processed using a large-language-model (LLM)–assisted extraction pipeline designed to identify the occupations of protagonists and their romantic partners. The LLM parsed book summaries and blurbs, as well as its own training knowledge, to extract candidate occupations, which were then standardized and manually validated through human review. This semi-automated procedure balances the scalability of computational text analysis with the interpretive accuracy of human coding, yielding a high-quality dataset suitable for sociological analysis.

The resulting dataset provides the first large-scale quantification of occupational portrayals across contemporary popular fiction. It enables comparison of which professions dominate cultural imagination, how occupational distributions differ by genre, and how these fictional patterns align or diverge from real-world labor market data. By examining the occupational representation prevalence in fiction, the dataset sheds light on the symbolic hierarchies of work as they are reproduced through mass-market storytelling.

1.5.3 Shared Methodological Principles

Although the three research components of this thesis differ in data source and substantive focus, they are unified by a common methodological orientation. Each project applies computational methods to uncover *relational structure*—

that is, patterns of similarity and association that cannot be captured through categorical or aggregate approaches.

A key methodological principle is the use of embeddings and clustering to model these relationships. In the labor market analyses (Chapters 3 and 5), occupations are embedded in a vector space derived from observed transitions between jobs, allowing similarity to be measured as continuous distance rather than binary equivalence. In the cultural analysis (Chapter 4), a parallel logic applies: occupations are extracted from text and compared against empirical baselines to quantify representation and prestige. In both domains, the goal is to replace rigid taxonomies with flexible, empirically grounded representations that capture the latent organization of social and cultural systems.

All projects also share a commitment to scale and interpretability. By drawing on millions of registry records or over a thousand novels, the analyses extend beyond the small-N case studies typical of both classical sociology and cultural analysis. Together, these methodological principles define the contribution of this dissertation not simply as the application of new tools, but as the demonstration of how computational approaches can be theoretically meaningful, empirically rigorous, and sociologically interpretable.

1.6 Contributions of the Dissertation

1.6.1 Empirical Contributions

Empirically, this dissertation advances the study of occupations by generating and analyzing new large-scale datasets that illuminate both the structural and symbolic dimensions of work. Across the three projects, the thesis provides evidence on how occupations are related and represented, offering a more comprehensive picture of how labor markets and cultural narratives jointly organize social life.

Chapter 3 provides an empirical mapping of the Danish labor market into occupation communities derived from population-wide registry data. This data-driven representation captures the latent structure of occupational relations as revealed by real mobility patterns, offering a fine-grained portrait of how jobs

cluster in practice. The analysis further characterizes these communities by their demographic and economic profiles and traces how educational cohorts move through them over time. Finally, it identifies which parts of the labor market were most disrupted by the COVID-19 pandemic, revealing enduring inequalities in resilience and recovery.

Chapter 4 adds an entirely different type of empirical evidence: a large-scale dataset of 1,181 bestselling novels in which protagonist and love-interest occupations were systematically identified using an LLM-assisted pipeline with human validation. This corpus enables quantitative analysis of how work is portrayed in contemporary mass culture, revealing strong genre patterns and gender asymmetries in the symbolic representation of labor. The findings demonstrate that certain archetypes—creative, romantic, or high-risk professions—dominate fiction, while the routine or care-oriented work sustaining most lives is largely invisible. This empirical contribution provides the first large-scale measurement of occupational representation in popular fiction.

Finally, the exploratory study on assortative marriage introduces evidence on partner matching using occupational distance rather than categorical identity. By applying embedding-based similarity measures to registry data on Danish couples, it reveals graded patterns of occupational homogamy that traditional same-job metrics obscure.

Taken together, these three components constitute a broad empirical contribution: they map the structure of work both as it exists in labor markets and as it is reproduced in cultural products.

1.6.2 Theoretical Contributions

Theoretically, this dissertation advances a relational and dual-aspect view of occupations as both structural positions within labor markets and symbolic resources within cultural narratives. Rather than treating occupations as fixed, atomistic categories, it conceives of them as dynamic and meaning-bearing units that link the economic and cultural dimensions of social life.

From a structural standpoint, the thesis models occupations by the mobility between them rather than elements of a predefined taxonomy. Chapter 3 shows

that occupational similarity can be inferred from observed transitions between jobs, a relational framing which complements traditional class schemas. The exploratory study on assortative marriage extends this logic, suggesting that the same relational configuration shaping career trajectories also informs patterns of partner choice.

From a symbolic standpoint, the thesis demonstrates that occupations also function as cultural signifiers. Chapter 4 shows how fiction selectively amplifies certain professions while omitting others, constructing symbolic hierarchies that diverge from those observed in labor-market data. This asymmetry underscores that social meaning cannot be inferred from economic position alone: occupations operate simultaneously as categories of work and as narrative devices.

Together, these perspectives integrate insights from stratification sociology, cultural sociology, and computational social science. By aligning embedding-based and text-based representations of work, the dissertation proposes a framework for studying occupations as both relational and representational phenomena—bridging the structures that organize labor with the stories that give it meaning.

1.6.3 Methodological Contributions

Methodologically, this dissertation demonstrates how computational tools can be harnessed to revisit and extend core social scientific questions about work, inequality, and meaning. Across the three projects, it develops and applies techniques that move beyond the use of predefined categories toward empirically derived, relational representations of occupations. The contribution lies in adapting methods in new contexts, and showing how they can be theoretically grounded and empirically validated.

The embedding-based framework of Chapter 3 scales classical sociological analyses of mobility to population-level data while preserving interpretability through community detection and demographic profiling. It provides a flexible foundation for comparative research across time and space, and a methodological template for future studies of complex social systems.

Chapter 4 advances computational methods in cultural sociology by demonstrating how large-language models (LLMs) can be integrated with human validation to extract structured sociological information about cultural products. The LLM-assisted pipeline used to identify protagonist and love-interest occupations in over a thousand bestselling novels exemplifies a hybrid approach: scalable, yet sufficiently validated to maintain interpretive rigor.

The exploratory study on assortative marriage extends the embedding framework developed in Chapter 3 into the measurement of social proximity. By operationalizing occupational similarity as a continuous distance rather than a categorical match, it refines a classical sociological construct—homogamy—and demonstrates how computational methods can yield more sensitive and realistic measures of relational patterns.

1.7 Thesis Structure

The remainder of this thesis is organized into six chapters.

Chapter 2 (*Methods*) provides a detailed account of the data sources, modeling techniques, and analytical procedures used across the three empirical studies. It describes the structure and linkage of the Danish administrative registries, the preprocessing of occupational sequences, and the training and validation of the embedding and clustering models used to identify occupation communities. It also outlines the construction of the fiction corpus and the LLM-assisted extraction and validation of occupational data, as well as the approach used to calculate embedding-based occupational distances in the assortative marriage study. Finally, the chapter discusses considerations of interpretability, reproducibility, and ethical data handling.

Chapter 3 (*Manuscript 1: Data-driven Coarse-graining of Occupations*) presents the first study. Using population-wide Danish registry data from 2011–2022, it develops an embedding-based approach to identify 311 empirically derived occupation communities. The analysis documents demographic, educational, and wage variations across these communities and demonstrates their utility for understanding labor market resilience.

Chapter 4 (*Manuscript 2: Occupations in Contemporary Popular Fiction*) presents the second study. It compiles and analyzes a corpus of 1,181 best-selling novels from 2023–2024 to examine how occupations are represented in contemporary fiction. Using an LLM-assisted extraction and validation pipeline, the chapter compares fictional portrayals of work with real-world occupational distributions, highlighting the cultural construction of prestige and desirability.

Chapter 5 (*Exploratory Study: Assortative Marriage by Occupational Distance*) extends the methodological framework of the first manuscript to the study of homogamy. By representing occupations as embedding vectors and comparing real couples to random baselines, it introduces a distance-based measure of occupational similarity that refines classical models of assortative marriage.

Chapter 6 (*Discussion*) integrates the findings from the three studies to draw out their broader implications. It situates the results within existing theories of class, mobility, and cultural representation, highlighting how data-driven approaches can reveal new insights on occupational structure and meaning. It also reflects on the methodological implications of using computational methods in sociological research, assessing their potential and limitations.

Chapter 7 (*Conclusion*) summarizes the overall contributions of the thesis and outlines directions for future research. It closes by considering how emerging data sources and analytical techniques may further expand the capacity of computational social science to address classical sociological questions about work, value, and social organization.

2 Materials and Methods

2.1 Overview

This chapter outlines the data sources, computational procedures, and analytical frameworks used across the three empirical components of the dissertation. While each project is presented as a stand-alone study in the subsequent chapters, they share a common methodological foundation: the use of large-scale data to model social and cultural structure through relational similarity. Together, they combine population-level administrative registries with large-scale text corpora, applying embedding, clustering, and comparative modeling techniques to uncover patterns that traditional categorical approaches cannot capture.

The sections that follow provide a unified description of the materials and methods underlying all projects, with additional technical detail beyond what is included in the manuscripts themselves. They cover the construction and preprocessing of Danish registry data, the creation of the fiction corpus and its LLM-assisted occupational coding, and the development of embedding-based measures of occupational similarity. Where appropriate, project-specific adaptations are discussed in greater depth in the corresponding empirical chapters. The goal of this chapter is to make the methodological foundation of the dissertation transparent, reproducible, and comparable.

2.2 Chapter 3: Data-driven Coarse-graining of Occupations

2.2.1 Data Source: Danish Labour Market Account

The primary data source for this project was the AMRUN (Labour Market Account) registry maintained by Statistics Denmark (Stender *et al.*, 2015). Established in 2008, this registry provides monthly labor market information for the entire Danish population, regardless of employment status. All residents are included—children, students, and retirees as well as active workers—and individuals may have multiple concurrent records within a month, each corresponding to a distinct labor market tie. These ties encompass a range of statuses, including employment, self-employment, unemployment benefits, educational enrollment, and other forms of activity outside the labor force. Each record contains detailed information such as start and end dates, income, employer or institutional identifier, and sectoral affiliation.

The variables used in this project are listed below (in alphabetical order):

- ALDER_AMR: Age at the end of the month.
- ARB_HOVED_BRA_DB07: Industry code of employer or business.
- BREDT_LOEN_BELOEB: Monthly salary in DKK, including non-wage benefits and pension contributions.
- DISCO_KODE: Occupational classification code.
- FOED_DAG: Date of birth.
- FRAVAER_BESK_KODE: Indicator for temporary absence from work (e.g., sick leave, parental leave).
- FRA_DATO: Start date of the labor market spell.
- I_BEFOLKNINGEN_KODE: Indicator for residency status (in Denmark vs. abroad).
- KOEN: Sex.

- PNR: Unique personal identifier enabling linkage across Statistics Denmark's registries.

- PRIMAER_STATUS_KODE: Indicator for an individual's primary labor market status in a given month.

- SOC_STATUS_KODE: Socioeconomic status category indicating labor market tie.[1]

- TILSTAND_GRAD_AMR: Ratio of hours worked during the spell to a standard full-time workload.

- TILSTAND_LAENGDE_MDR: Proportion of the month for which the labor market status applied.

- TIL_DATO: End date of the labor market spell.

- UDD_BESK_KODE: Indicator for simultaneous employment and enrollment in education.

- VERSION_NR: Registry version number.

Although the registry extends back to 2008, the analysis focused on the years 2011–2022. This restriction was necessary because the key occupational variable, DISCO_KODE, transitioned to a new encoding format in January 2010, and the months surrounding the change contain substantial coding inconsistencies. The analytic sample included all individuals aged 15–65 who were recorded as residing in Denmark and had at least one working state (employment, self-employment, or co-working spouse status) during the study period. After applying these filters, the dataset comprised 4,373,896 individuals.

Occupations were coded using DISCO_KODE, a Danish adaptation of the International Standard Classification of Occupations (ISCO-08) (Statistics Denmark, 2011; International Labour Office, 2023), and industries were coded according to ARB_HOVED_BRA_DB07, which corresponds to the NACE Rev. 2 classifica-

[1] Note that this variable does not correspond to the regular meaning of the term "socioeconomic status", but rather to labor market states, e.g., "wage earner," "self-employed," "unemployed," etc.

tion (Eurostat, 2008). In total, the dataset included 784 distinct occupation codes and 736 industry codes.

2.2.2 Supplementary Registry: ELEV2 (Education)

To contextualize occupational patterns within educational background, I supplemented the Labour Market Account with the ELEV2 registry (Sørensen, 2024), which records educational enrollments and completions for all individuals in Denmark. The registry covers the period 1991–2014 and includes information on institution, program code, level of qualification (e.g., vocational, bachelor's, master's), and field of study according to the Danish Education Classification (UDDA). Each individual is identified through the same anonymized personal identifier (PNR) used across Statistics Denmark's systems, allowing direct linkage to monthly labor market histories in AMRUN.

For this project, the ELEV2 data were used to examine how graduates from different educational programs dispersed through the empirically derived occupation communities over time. Educational completion year and field were used to construct graduate cohorts, which were then followed longitudinally in the labor market data to assess variation in early-career mobility and clustering patterns.

2.2.3 Preprocessing and Labeling

Prior to constructing occupational sequences, the registry data were cleaned and standardized to ensure consistent labeling across employment types and firm sizes. While the Labour Market Account provides near-complete coverage of the Danish workforce, some missingness remained in occupational labels (DISCO_KODE). These gaps arose because employers with fewer than ten employees are not required to report occupation codes. Approximately 9% of wage earners therefore lack this variable, though other information—particularly industry codes and employment type—was available for most cases.

For individuals without a valid DISCO_KODE, the employer's industry code (ARB_HOVED_BRA_DB07) was substituted, as it was available in roughly 95%

of affected records. For self-employed individuals and co-working spouses, occupation codes are systematically absent but industry codes are typically present (97% coverage). These entries were therefore labeled by combining the available industry code with an indicator of employment type (wage earner, self-employed, or co-working spouse). This approach preserves both economic context and role distinction while maintaining compatibility with the occupational embedding procedure.

The resulting labeling scheme produced 2,698 distinct occupational identifiers, defined as unique combinations of occupation or industry code and employment type.

2.2.4 Embedding Construction (word2vec)

To capture the empirical structure of career mobility, a word2vec model (Mikolov *et al.*, 2013) was trained on sequences of occupations derived from individual labor market histories. Each person's employment record was transformed into a "sentence" representing their chronological transitions between occupations, where each "word" or token corresponds to an occupation label. To reduce redundancy and emphasize transitions rather than tenure, only the first month of each continuous employment streak was retained as a token.

When individuals held multiple concurrent jobs within the same month, job order was randomized to avoid systematic bias stemming from the registry's listing order. Extended unemployment spells (six months or longer) were included as a dedicated token to capture exits from the labor market, and education periods were included when the individual reported at least ten hours of educational activity within a month. All educational episodes were represented by a single shared token, allowing the model to recognize education as a general career state rather than a distinct occupation.

All individual sequences were concatenated into a single corpus, representing the population's aggregate occupational mobility. The word2vec model was trained on 80% of the sequences using the Skip-gram architecture for 95 epochs, with a vector dimensionality of 100, a context window of seven tokens, a minimum word count of one (retaining rare occupations), and a down-sampling rate of 0.1. These parameters were selected based on minimizing

validation loss in a proxy prediction task (predicting masked occupations from context on the remaining 20% of sequences). Validation plots are presented in Appendix A.1.

The resulting embedding assigns each occupation a continuous vector representation, positioning occupations closer together if they tend to co-occur in individuals' career trajectories. This contrasts with transformer-based models, where the representation of a token changes based on context from neighboring tokens.

This data-driven representation provides the foundation for subsequent clustering and analysis.

2.2.5 Clustering and Community Detection

To construct an empirical coarse-graining of the Danish labor market, I identified occupational communities directly from the embedding space learned by the word2vec model. This procedure captures patterns of career mobility and groups structurally similar occupations.

Projection and HDBSCAN Clustering
Because the word2vec embeddings are high-dimensional and thus prone to the curse of dimensionality, I first scaled and standardized the embeddings before reducing their dimensionality using t-SNE, which preserves local structure and facilitates detection of community-like groupings in a lower-dimensional space (Thrun and Ultsch, 2021). The embeddings were projected into two dimensions with a perplexity parameter of five.

I then applied HDBSCAN (McInnes *et al.*, 2017), a hierarchical density-based clustering algorithm chosen for its robustness to noise and its ability to identify clusters of varying density. HDBSCAN was run with a minimum cluster size of 3 to restrict focus to non-trivial communities, and a minimum sample size of 3, using the "excess of mass" (`eom`) criterion for cluster selection. The resulting partition included 311 occupation communities and 355 occupations that were classified as noise (unclustered). The distribution of community sizes was relatively uniform, with the largest cluster containing approximately 30 occupations.

This projection-based clustering approach provides a data-driven coarse-graining of the labor market: occupations that commonly co-occur in individual career sequences are grouped together into empirically coherent labor market communities. These communities serve as the primary analytical units in subsequent analyses.

Alternative Community Detection Approaches

To evaluate whether the embedding-based method provides a meaningful representation of occupational structure, I compared it with several alternative network-based clustering approaches, reflecting the methods most commonly seen in the relevant literature (e.g., Cheng and Park, 2020; Rio-Chanona *et al.*, 2020; Toubøl and Larsen, 2017). Even before clustering, network- and embedding-based approaches have trade-offs—in particular, representing transition weights is more natural using networks, while representing worker histories is more natural using embeddings.[2]

First, I constructed a directed, weighted network of job transitions, where nodes represent occupations and edges represent empirical transitions between them. Self-directed edges denote continued employment in the same occupation.

Edge weights were equal to the normalized number of hours transitioned between occupation states. In cases where individuals held multiple concurrent positions, each job's monthly hours were scaled to sum to one, and fractional changes in this distribution were used to compute weighted transitions. This procedure accommodates transitions involving multiple income sources, including part-time or self-employment, while normalizing each individual's total monthly weight to one. Non-employment states such as unemployment or education were removed as nodes to avoid spurious linkages caused by their ubiquity as transition intermediates.

The resulting Danish occupational mobility network for 2011–2022 contains 2,698 nodes and 746,020 weighted edges, of which 2,696 are self-directed. The network is highly dense and connected, forming a "hairball" structure even after applying backbone filtering techniques such as minimum-weight

[2]Notwithstanding network clustering methods that take longer walks across the network into account (e.g., Persson *et al.*, 2016) which solve this issue, but fail to scale to large data with longer paths.

thresholds, the disparity filter (Serrano *et al.*, 2009), or noise correction (Coscia and Neffke, 2007). This density contrasts with sparser U.S.-based networks (e.g. Escobari *et al.*, 2021; Rio-Chanona *et al.*, 2020) and likely reflects the higher granularity and full-population coverage of the Danish registries. Since backbone filtering techniques resulted in information loss with no uncovering of latent structure, I used the raw network with no filtering for subsequent analyses unless otherwise indicated.

To detect communities within this network, I tested two complementary methods: Infomap (Rosvall and Bergstrom, 2008) and adaptive link clustering (Ahn *et al.*, 2010). Infomap partitions nodes by minimizing the description length of random walks across the network, identifying dense subgraphs that reflect mobility modules. Link clustering instead groups edges rather than nodes, allowing community membership of nodes to overlap—appropriate for heterogeneous data in which nodes may belong to multiple functional domains.

Evaluation and Comparison of Methods

To assess clustering quality across methods, I adapted four evaluation metrics from Ahn *et al.* (2010): *community coverage*, *overlap coverage*, *community quality*, and *overlap quality*. Community coverage measures the proportion of occupations placed in nontrivial communities (those with at least three members), while overlap coverage captures the average number of communities per node. Community quality quantifies the degree to which empirically derived communities align with external metadata, and overlap quality compares how many communities a node is expected to belong to versus how many it actually does belong to.

To evaluate community quality, I used official textual descriptions of ISCO and NACE codes, which were embedded using a pre-trained BERT model, (Reimers and Gurevych, 2019) to produce a similarity matrix based on cosine distance—independent of the empirical transition data. This BERT variant was chosen as it was trained to derive embeddings from full English-language sentences and paragraphs. As a benchmark for theoretical ”maximum” enrichment, I compared all methods against a baseline clustering that grouped occupations by ISCO and NACE sub-major categories, which yielded a reference quality score of 2.87. For overlap quality, I assumed that occupations held by a greater number of unique individuals should appear in a greater number of

communities, reflecting their broader reach within the labor market. Table 2.1 summarizes results across methods.

Table 2.1.. Quality metrics for different coarse-graining methods.

Community detection method	Community quality	Community coverage	Overlap quality	Overlap coverage
Infomap	1.42	0.991	N/A	N/A
Recursive Infomap	1.54	0.989	N/A	N/A
HDBSCAN	1.90	0.868	N/A	N/A
Link clustering	1.39	0.593	0.184	3.08×10^{-2}
Link clustering, min. 10 individuals	2.28	0.348	0.118	0.168

Infomap identified 51 modules (31 nontrivial) with 75% relative codelength savings, though the results were dominated by two very large clusters encompassing more than half of all occupations. Applying Infomap once again on the two largest clusters produced 60 total modules (38 nontrivial), but the size distribution remained heavily skewed. This skew likely reflects the density structure of the underlying network—a densely connected core surrounded by a sparser periphery—which differs from the structural assumptions of the Infomap algorithm.

Link clustering produced fractional membership information for nodes across multiple clusters, often with extremely low membership weights. To obtain interpretable communities, I filtered for cases where a node's membership in a given community exceeded 5%, yielding 83 nontrivial communities. The resulting clusters exhibited comparatively low community quality. To address the fact that link clustering does not incorporate edge weights, I also tested performance by running the algorithm on a filtered version of the network where edges representing fewer than ten individuals were removed; this produced 452 nontrivial communities and improved overall quality, albeit at the expense of community coverage.

By contrast, projection-based clustering of the embedding space using HDBSCAN achieved the most balanced performance across all evaluation metrics, and required no ad hoc tuning such as weight thresholds or recursive runs. It produced 311 interpretable, non-overlapping communities with a relatively even size distribution, avoiding the dominance of a few large clusters characteristic of network-based methods. For these reasons, I adopted the embedding-based approach as the primary method for subsequent analyses.

2.2.6 Ethical Considerations

All analyses were conducted within the secure research environment of Statistics Denmark, which provides controlled access to administrative microdata under strict confidentiality protocols. Researchers cannot access identifiable information; all datasets are anonymized before access and remain stored on Statistics Denmark's servers. Data extraction, copying, or external storage of raw records is prohibited, and only aggregated, non-disclosive outputs may be exported following approval by Statistics Denmark.

The study involved no contact with individuals, and all analyses were performed on pseudonymized population-level data. No attempt was made to re-identify individuals or entities, and all reported results were subject to Statistics Denmark's disclosure control checks before release.

2.3 Chapter 4: Occupations in Contemporary Popular Fiction

2.3.1 Corpus Construction

To analyze how occupations are represented in contemporary popular fiction, I compiled a corpus of bestselling novels published or widely sold in the U.S. 2023–2024.

The initial list of titles was assembled by scraping multiple major bestseller sources on Publishers Marketplace: the ABA Indiebound Hardcover and Paperback Fiction lists, Amazon's Combined Print/eBook Fiction rankings, and the New York Times Hardcover and Paperback Fiction charts. These lists collectively represent the U.S. trade fiction market, spanning independent bookstores, online retail, and traditional publishing houses. All available lists published in the study period, generally of weekly frequency, were scraped, and duplicate entries were merged by title and author strings.

To add basic metadata, I queried the Google Books API for each title, retrieving book descriptions and coarse genre tags. The Google Books genre labels

were used primarily to distinguish novels from other literary forms such as nonfiction, poetry, or graphic novels. When classification tags were ambiguous or inconsistent with the synopsis, I cross-checked the title's dominant genre using Goodreads metadata. Discrepancies were resolved manually through inspection of publisher blurbs and first-page previews.

The unit of analysis is the unique novel title. Duplicate editions (e.g., hardcover and paperback) were consolidated. Titles that did not meet the inclusion criteria—including anthologies, short-story collections, graphic novels, and poetry volumes—were removed from the dataset. In total, 180 titles were excluded during preprocessing.[3] A small number of additional non-novel works that escaped initial filtering were identified and removed during later validation.

After cleaning, the final corpus contained 1,181 unique bestselling novels. For approximately half of these ($n = 533$), full narrative summaries were available and were appended to the metadata via automated scraping from SuperSummary, a website which provides novel summaries and study guides. Each record in the final dataset includes standardized title and author fields, book blurbs, and (where available) plot summary text.

The complete dataset and accompanying preprocessing scripts are available in a public GitHub repository at `https://github.com/ella-clement/fiction-professions`.

2.3.2 Automated Extraction of Professions and Metadata

To extract structured information about narrative and occupational features, I used a large language model (LLM) to parse book metadata and plot summaries into a consistent coding schema. The extraction targeted four key dimensions: (1) genre classification, (2) identification of main protagonists and love interests, and (3) their stated or implied professions.

[3] See `https://github.com/ella-clement/fiction-professions/blob/main/data/removed_titles.xlsx` for the full list.

All extractions were conducted using GPT-4o accessed through the OpenAI API (OpenAI, 2024). The model was prompted in a multi-stage pipeline that combined open-ended extraction with targeted follow-up queries to improve precision and reduce hallucinations (see Appendix B.1 for the full prompt template). All prompts were run with a temperature parameter of 0.3, a relatively low value suitable for reducing hallucinations and producing more reliable, deterministic output.

In the first stage, the model received the book title and author, and was asked to identify the central characters, their relationships, and their professions or primary roles in the narrative. In the second stage, this query was repeated, this time with synopses (from Google Books and/or SuperSummary), to allow the LLM to reason based on extra knowledge of the work, which often resulted in higher accuracy. In the third stage, the outputs were merged by selecting the result for each field from the first stage, unless the field was marked `Unknown`, and otherwise selecting the result from the second stage.

When a novel featured multiple protagonists or an ensemble cast, all central figures were coded. Occupation fields were recorded as ”None” when no employment or professional identity was specified, and categories were introduced for common non-employment categories (e.g., child, student, retired).

Genre classification was initially performed automatically via GPT-4o and then standardized manually. Synonymous or overlapping subgenres were harmonized (e.g., ”fantasy romance” and ”romantasy”), and narrower categories (e.g., ”dystopian,” ”paranormal”) were subsumed under broader umbrella genres (”science fiction,” ”fantasy”). I cross-referenced Goodreads metadata and reader tags to confirm genre assignment for all titles. The final genre taxonomy included ten high-level categories: contemporary, fantasy, fantasy romance, historical, horror, literary, mystery, romance, science fiction, and thriller.

For the subset of romance and fantasy romance novels, I additionally hand-coded the gender of protagonists and love interests as ”man” or ”woman,” based on the gender presentation of characters within the text and associated paratexts (e.g., blurbs and publisher descriptions). This binary coding reflects the universal treatment of gender in these genres and facilitates comparison

of gendered patterns in occupational desirability and narrative role. Cases featuring same-gender couples were coded explicitly as such.

2.3.3 Validation and Accuracy Assessment

To evaluate the reliability of automated coding, I implemented a human validation procedure based on a random 10% subsample of the corpus ($n = 117$ novels). Half of this sample ($n = 59$) served as a development set used iteratively to refine the prompting and post-processing pipeline, while the remaining half ($n = 58$) was held out as an independent validation set.

Each book in the validation sample was manually reviewed to extract the same variables as in the automated coding: genre, protagonist name(s) and profession(s), and, where applicable, love interest(s) and their professions. In cases of ambiguity, I consulted the book's blurb and, if necessary, read excerpts or the full content to confirm key details. For romance and fantasy romance titles, gender of the protagonists and love interests was also verified manually.

Accuracy metrics are reported in Table 4.1. Automated extraction achieved 92–95% accuracy for protagonist identification and 80–83% for occupation extraction, both within or above accepted intercoder reliability thresholds in manual content analysis (typically $\kappa \geq 0.80$; see Neuendorf, 2017). Accuracy was highest for books with straightforward narrative structures and detailed plot summaries, and lowest for multi-perspective or ensemble narratives with ambiguous occupational roles. The majority of errors stemmed not from incorrect assignments but from the model returning "unknown," accounting for approximately 77% of all profession-related inaccuracies.

The validated dataset was then reintegrated into the main corpus by replacing the corresponding automated outputs for the reviewed novels with hand-coded values. This ensured that approximately 10% of the total corpus reflects verified human coding while the remaining 90% retains automated extractions.

2.3.4 External Comparison Datasets

To contextualize the occupational patterns observed in fiction, I incorporated three types of external benchmark data: (1) U.S. occupational prevalence and wage distributions, (2) survey-based prestige scores, and (3) measures of occupational desirability in romantic partners. These sources allowed comparison between fictional portrayals of work and real-world occupational structures, economic valuation, and social or romantic appeal.

Each dataset required harmonization with the fictional corpus, including alignment of occupational taxonomies, aggregation to comparable resolution, and handling of categories absent from one or more datasets. All external data were treated descriptively rather than inferentially: the goal was to situate cultural representations of work relative to structural and attitudinal baselines, not to establish causal relationships. The subsections that follow describe each external dataset in turn, along with its source, harmonization procedure, and analytical role.

Real-world Prevalence and Wages

To assess how occupational representations in fiction align with real-world labor market structure, I compared the frequency of fictional occupations to U.S. employment and wage data. The primary source was the Bureau of Labor Statistics' Occupational Employment and Wage Statistics (OEWS) dataset (U.S. Bureau of Labor Statistics, 2024), which provides national counts and average hourly wages by detailed occupation.

To harmonize the fictional dataset with the OEWS taxonomy, I first restricted the fictional sample only to novels in genres that are generally set in contemporary times in a real-world location: contemporary fiction, horror, literary fiction, mystery, romance, and thrillers. Next, I aggregated or merged occupational categories where necessary—for example, combining "novelist" and "poet" under "writer," or consolidating all postsecondary teaching roles when subject specialization could not be identified in fiction. Mapping proceeded at the most specific level possible, but in many cases, broader alignment was required to ensure comparability between fictional and real-world occupations. As a consequence, some roles may overlap—for example, "doctor" and "psychiatrist."

Certain fictional professions had no direct real-world equivalent in the OEWS dataset. Some represented non-employment states (e.g., student, retired), while others referred to rare or illicit occupations (e.g., bounty hunter, crime boss). For such cases, I used secondary online sources, including industry reports, professional associations, and publicly available workforce estimates, to derive approximate U.S. prevalence figures. These values should be interpreted as order-of-magnitude estimates, which is sufficient for comparative purposes since fictional and real-world frequencies were analyzed on a log scale. A summary of these secondary sources is provided in Table B.1.

Occupations for which no plausible real-world analog or estimate could be established were excluded from quantitative comparison, as were real-world occupations that never appeared in the fictional corpus. For wage data, I computed weighted averages when fictional categories encompassed multiple OEWS occupations. Missing wage estimates were supplemented using publicly available data from Zippia (*Zippia* 2025) and Glassdoor (*Glassdoor* 2025a). In all, there were 178 fictional roles which could be matched to a real-world counterpart, of which 156 had available salary information.

Prestige Scores

To examine how cultural representation relates to social valuation, I compared fictional occupational frequencies with survey-based prestige scores. The prestige data were drawn from Hughes *et al.* (2024), a large-scale online survey conducted via Amazon Mechanical Turk (MTurk) in 2015–2016. The study collected ratings from 3,076 respondents who evaluated the perceived prestige of 1,029 distinct occupations on a 0–100 scale. Although MTurk samples are not fully representative of the U.S. population, they are widely used in social science research (Berinsky *et al.*, 2012), and the resulting prestige rankings are highly correlated with long-established scales such as those reported in Treiman (1977).

To harmonize this dataset with the fictional corpus, I manually mapped occupation labels across the two sources, resolving minor lexical discrepancies and consolidating semantically equivalent titles under shared categories. This process yielded a reduced set of overlapping occupations with both reliable real-world prevalence estimates and prestige scores (143 roles). Analyses based on this intersection therefore focus on how fiction emphasizes, omits, or distorts occupations across the prestige hierarchy.

Romantic Partner Preferences

To assess how occupational portrayals in romance fiction align with real-world perceptions of desirability, I compared the occupations of fictional protagonists and love interests to three datasets capturing occupational attractiveness in romantic contexts. Two datasets were drawn from Tinder-based samples that report match rates by listed profession (Binder, 2016; Education Connection, 2022), while a third was derived from a Zippia survey ranking the most attractive occupations in potential partners (Morris, 2021). Together, these sources provide a pragmatic, if imperfect, proxy for occupational desirability across genders.

Each dataset carries distinct methodological limitations. The Tinder samples reflect only users who voluntarily disclose occupational information and are influenced by the app's demographic composition, cultural norms, and gender ratios. The Zippia survey, by contrast, was conducted through an opt-in online panel and included approximately 250 common occupations but reported minimal demographic details. All three sources measure heterosexual desirability exclusively and do not include same-gender preferences. None are representative in a statistical sense, yet collectively they capture patterns that have circulated widely in media coverage and public discourse, making them useful indicators nonetheless.

Because the three datasets cover different sets of occupations, rank correlation analysis was not meaningful. Instead, I focused on the six top-ranked professions from each source, corresponding to the shortest of the lists. Occupational labels were retained as reported in the original sources, as full underlying occupation lists were unavailable. The analysis therefore provides an indicative rather than exhaustive comparison between fictional romantic archetypes and real-world perceptions of occupational appeal.

2.3.5 Ethical and Practical Considerations

The analyses in this project relied exclusively on publicly available cultural products—bestseller lists, book summaries, and reader-generated metadata—supplemented with automated text processing and LLM-assisted extraction. No copyrighted text was redistributed, and all model outputs were limited to structured metadata (e.g., character roles, professions, and genres) rather

than verbatim reproduction of source material. The project therefore falls within the bounds of fair research use and does not involve human subjects or identifiable personal information.

All LLM-assisted analyses were conducted transparently, with full prompt template documented in Appendix B.1. While large language models introduce potential biases—reflecting cultural and linguistic asymmetries present in their training data—these were mitigated through manual validation. The validation exercise described above provides a quantitative estimate of accuracy and helps ensure that downstream analyses are not dominated by model-specific artifacts.

Attention was also paid to reproducibility. While exploratory work indicated that LLMs connected to the internet, for example ChatGPT-4o, had higher performance, using these would make it impossible for other researchers to recreate the dataset using the same procedure. Additionally, it would have lowered the ability to choose a model and benchmark performance across different LLMs—hence the decision to query GPT-4o via API and provide the model with additional context.

Cultural datasets themselves are not neutral: the bestseller lists used represent the U.S. market, established publishers, and genres with high commercial visibility. Similarly, supplementary data such as online surveys and Tinder samples carry demographic and cultural biases that shape their apparent "real-world" benchmarks. Throughout, I treat these sources as reflections of contemporary cultural circulation rather than as representative samples of global literary or social patterns.

All code, extraction templates, and processed data (excluding copyrighted materials) are available in the accompanying GitHub repository at `https://github.com/ella-clement/fiction-professions`.

2.4 Chapter 5: Assortative Marriage by Occupational Distance

2.4.1 Data Sources

This exploratory project linked individual-level labor market information from the Danish Labour Market Account (AMRUN) with marital data from one of Statistics Denmark's civil registries (CIVST2022). The goal was to examine occupational similarity between spouses across the Danish workforce using both traditional categorical and embedding-based measures.

The AMRUN registry provides monthly records for all Danish residents, including employment status (SOC_STATUS_KODE) and occupation (DISCO_KODE). As described in Section 2.2.1, it covers the full working-age population and contains detailed employment histories for the years 2011–2022. Occupations are coded using the Danish six-digit DISCO scheme, where the first four digits correspond to ISCO-08 codes (International Labour Office, 2023) and the last two digits provide additional refinements. For each individual and month, I defined the primary occupation as that associated with the greatest number of hours worked.

The CIVST2022 registry records all legally recognized marriages and civil partnerships in Denmark as of the end of 2022, including both partners' anonymized identifiers, the date of marriage formation (CIV_VFRA), and, when applicable, dissolution (CIV_VTIL). Couples were included in the analytic sample if both partners had at least one observed wage-earning occupation during the study period. Same-sex marriages were excluded for comparability with prior research on assortative marriage and because occupational distributions and matching dynamics may differ in ways that merit separate study.[4]

Record linkage between AMRUN and CIVST2022 was performed within Statistics Denmark's secure research environment using the unique personal identifier (PNR), which is consistent across all national administrative registries. The resulting dataset included all opposite-sex couples who entered a legal union between 2011 and 2022 and for whom both spouses have valid occupational data. After applying these criteria, the analytic sample consisted of 331,371 couples.

[4] See, e.g., Schwartz and Graf (2009) for discussions of gender- and sexuality-specific patterns in assortative matching.

To align employment histories with the timing of union formation, I indexed event time in months relative to the marriage date, $\tau \in \{-143, \ldots, +143\}$, with $\tau = 0$ denoting the month of marriage. This structure allowed for analysis of partners' occupational trajectories before and after union formation.

2.4.2 Operationalizing Occupational Distance

To quantify occupational similarity between spouses, I used three complementary measures that ranged from categorical to continuous, drawing on both formal classifications and empirically learned embeddings.

First, I replicated the conventional measure of occupational homogamy by coding couples as similar if both partners shared the same six-digit occupation code (`DISCO_KODE`); otherwise, the value was zero. Occupation codes are missing for individuals outside wage employment, in which case similarity was also set to zero. This binary indicator serves as a benchmark for comparison with continuous measures.

Second, to capture graded semantic proximity within the ISCO-08 taxonomy, I embedded the official textual descriptions of occupations using a pre-trained BERT model (Reimers and Gurevych, 2019) as described in Section 2.2.5. Each occupation description was represented as a sentence vector, and the similarity between the partners' occupations at time τ is defined as the cosine similarity of their respective vectors. Higher values indicate greater overlap in the skill and task content of the occupations as defined in the ISCO framework.

Finally, I operationalized occupational distance using the empirical embeddings from Chapter 3. Each occupation was represented by a 100-dimensional word2vec vector trained on observed Danish job transitions from 2011–2022 (context window 7, min count 1, negative sampling 0.1, 95 epochs). Cosine similarity between these vectors then reflects similarity of the two occupations. As opposed to Chapter 3, I did not perform dimensionality reduction with t-SNE prior to distance computations, as this projection preserves local but not global structure.

To evaluate the extent of assortative marriage beyond what would occur by chance, I generated a counterfactual benchmark through random recombina-

tion. For each observed couple (i, j) married in month m and year y, I drew K pseudo-couples (i', j') from others marrying in (m, y) and computed similarity for the resulting pseudo-couples. This preserved the marginal occupational and cohort distributions while disrupting the true pairing, providing a null distribution of expected similarity.

2.4.3 Analysis Strategy

The analyses proceeded in three steps, each designed to evaluate how occupational similarity relates to marriage formation and to compare continuous, embedding-based measures with categorical approaches.

Event-time comparison. To examine how similarity between partners evolves around the time of marriage, I computed the mean occupational similarity for real couples and for counterfactual pairings generated through the cohort-preserving random recombination procedure described above. The difference between these two averages, $\Delta(\tau)$, captures the degree of assortative marriage at each event time τ, measured in months relative to marriage. Ninety-five percent confidence intervals were obtained via cluster bootstrapping (1,000 replicates), resampling at the couple level and repeating the random recombination step with a new random seed in each replicate.

Nonparametric enrichment analysis. Next, to visualize how the probability of marriage varies across the full range of occupational similarity, I partitioned all potential dyads into equal-width bins of embedding similarity. For each bin b, I computed the enrichment ratio

$$E(b) = \frac{P(\text{married} \mid \text{similarity bin } b)}{P(\text{married overall})},$$

which indicates whether marriage occurs more or less frequently among dyads with a given level of occupational proximity. Values above one reflect over-representation of marriages among similar occupations, while values below one indicate avoidance. This approach provides an intuitive, model-free representation of the relationship between occupational distance and partner formation.

Model-based estimation. I estimated a set of matched logistic choice models to quantify how occupational similarity predicts marriage formation. For each observed marriage, I constructed a *matched set* consisting of the real couple and $K = 10$ counterfactual male–female pairings drawn randomly from all individuals who married in the same calendar month. Each matched set forms a single choice stratum, within which only one dyad is observed ($Y = 1$) and the others are unobserved ($Y = 0$). I then estimated conditional logistic regressions with three specifications:

$$\begin{aligned}
\text{(M1)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_1\, \text{SameOcc}_{ij}, \\
\text{(M2)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_2\, \text{EmbedSim}^*_{ij} + \sum_{r=1}^{2} \gamma_r\, \text{RCS}_r(\text{EmbedSim}^*_{ij}), \\
\text{(M3)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_1\, \text{SameOcc}_{ij} + \beta_2\, \text{EmbedSim}^*_{ij} \\
&\quad + \sum_{r=1}^{2} \gamma_r\, \text{RCS}_r(\text{EmbedSim}^*_{ij}),
\end{aligned}$$

where Y_{ij} represents whether i and j form an observed couple, EmbedSim^*_{ij} is the standardized version of EmbedSim_{ij}, and $\text{RCS}(\text{EmbedSim}^*_{ij})$ is a restricted cubic spline basis (with $k = 4$ knots, yielding $k - 2 = 2$ spline terms). The spline knots were placed at empirical quantiles of EmbedSim^*, and all spline components were evaluated on the standardized scale. The spline specification allows for smooth, flexible deviations from linearity while avoiding the global curvature implied by polynomial models, and less computational overhead than a generalized additive model.

Coefficients were estimated using conditional maximum likelihood, which conditions out stratum-specific intercepts and thus avoids the incidental parameters problem. Standard errors were clustered by marriage month (the unit governing the construction of counterfactual dyads). For interpretability, I report odds ratios for exact occupational matches as well as spline-based contrasts comparing predicted odds of marriage at, for example, the 25th vs. 75th percentile of the similarity distribution.

No additional demographic or regional covariates were included, as the purpose was not to isolate occupational similarity from other sorting mechanisms

but to quantify the aggregate degree to which shared labor market positioning corresponds to union formation.

2.4.4 Ethical Considerations

All analyses were conducted within Statistics Denmark's secure research environment, which ensures compliance with Danish data protection legislation and the General Data Protection Regulation (GDPR). Individual-level registry data are pseudonymized prior to researcher access: personal identifiers are replaced by encrypted keys, and all results must pass disclosure control before export. These procedures prevent identification of any individuals or couples while allowing linkage across registries through consistent pseudonyms.

Analyses involving family linkages, such as those using the civil registry to identify marriages, carry particular privacy sensitivities because they involve relational data that connect individuals to one another. To minimize disclosure risk, only the information necessary to establish marital ties (partner identifiers and dates of union formation) was used, and all derived variables were stored and analyzed in aggregated form. No attempts were made to infer or analyze cohabiting or non-marital relationships, which are only partially observable in the administrative data.

2.5 Cross-cutting Considerations

2.5.1 Reproducibility and Transparency

All projects in this dissertation were designed to be transparent and reproducible within the constraints of Danish data protection law. For registry-based analyses (Chapters 3 and 5), individual-level data cannot be shared or exported from Statistics Denmark's secure servers. However, all preprocessing, modeling, and analysis scripts are fully versioned and archived within the secure environment, enabling replication by other authorized researchers.

To facilitate reproducibility outside the secure enclave, I have released aggregated counterparts of all non-sensitive datasets. The GitHub and OSF repositories associated with this dissertation include:

- Anonymized community-level summaries of occupational communities, with aggregated data on community membership including salary percentiles, age, and gender distributions (relating to Chapter 3). See Appendix A.3 for further details.

- Complete, open-access code for cultural data collection, cleaning, and LLM-assisted extraction, as well as non-copyrighted datasets (relating to Chapter 4).

- Anonymized, binned similarities at month of marriage for each couple according to each of the three similarity metrics used in Chapter 5. See Appendix C.3 for further details.

2.5.2 Methodological Reflections

Across all three projects, the dissertation demonstrates how computational methods—specifically embeddings, clustering, and large language model (LLM)-assisted coding—can be adapted to address social scientific questions about work, inequality, and representation. These tools enable the detection of latent structures in large, complex datasets, offering empirical traction where traditional survey or case-based approaches face limits of scale. Yet they also introduce new epistemic and ethical challenges that warrant reflection.

Embedding models, as used in Chapter 3 and extended in Chapter 5, translate social processes into geometric space: occupations become points positioned according to empirical transition probabilities. This abstraction captures relational similarity in a way that conventional taxonomies cannot, revealing gradients of proximity and occupational neighborhoods. However, embeddings are inherently inductive and data-dependent. They may therefore fail to generalize across models or data settings.

Clustering techniques further distill these representations into discrete communities. The resulting ”coarse-grained” units are analytically useful for

summarizing labor market complexity, but they also impose categorical boundaries on what is, in reality, a continuous landscape. Choosing the right level of aggregation therefore requires a qualitative rather than purely statistical judgment: clusters must be interpretable in terms of occupational pathways, not just mathematically coherent.

For the cultural data in Chapter 4, LLM-assisted extraction allowed for large-scale, systematic coding of textual data that would otherwise be intractable. The hybrid pipeline—automation supplemented by targeted human validation—balances efficiency with accuracy. Yet the use of generative models raises concerns regarding bias and opacity: language models inherit the biases of their training data and may reproduce representational stereotypes. Transparency about prompting, validation, and error rates is therefore crucial for accurate use.

More broadly, these projects illustrate the trade-offs between interpretability and predictive accuracy, and between automation and human oversight. Computational methods can expand the scope of sociological inquiry but cannot fully replace theory-driven reasoning or critical interpretation. Their greatest strength lies in complementing, not substituting, human judgment, providing new empirical foundations for understanding the social structures and cultural narratives that shape work and identity.

3

Data-driven Coarse-graining of Occupations

This chapter is based on the manuscript:
Ella Clement, Andreas Bjerre-Nielsen, Niels Richard Hansen, Sune Lehmann. *Labor market embeddings reveal novel dynamics and structures in the Danish labor market*. Submitted to PLOS ONE, 2025.

Abstract

Understanding the structure and dynamics of labor markets is essential for analyzing economic inequality, resilience, and workforce development. However, traditional occupational classifications rely heavily on expert-defined taxonomies, which may overlook empirical patterns present in real-world dynamic career trajectories. This study constructs a data-driven coarse-graining of the Danish labor market based on occupational mobility for 4,373,896 Danish workers over 12 years. Our premise is that occupational properties such as wage inequality and demographic characteristics are expressed throughout career progressions rather than intrinsic characteristics of single jobs. Using comprehensive registry data on the Danish population (2011–2022), we identify 311 occupation communities. To identify communities, we used word2vec and a subsequent clustering step. Our communities vary significantly in age structure, gender balance, self-employment rates, and wage distributions. We show that this holistic classification uncovers new empirical patterns in labor market dynamics. In particular, we find that different educational programs produce markedly different degrees of occupational spread, which generally increase with time post graduation. Finally, we identify persistent labor market disruptions in a small number of communities following the COVID-19 lockdown, characterized by sharp exits and no full recovery in labor force share.

Our findings highlight the structural dynamics of the Danish labor market and underscore the value of empirically derived occupation communities for studying workforce evolution and inequality.

3.1 Introduction

Occupational mobility—how individuals transition from job to job—reveals general patterns in career trajectories and labor market structure (Blau and Duncan, 1967). Understanding these patterns is important for analyzing economic adaptability to labor market shocks and income inequality (Kalleberg, 2009; Cahuc *et al.*, 2014). While occupational mobility has been modeled previously, most existing work considers only the most recent job transition and often omits key labor market dynamics such as self-employment, part-time work, and longer-term career histories.

In this study, we aim to model labor mobility empirically in a way that remedies these gaps. We identify groupings of similar jobs—a coarse-graining of the labor market—based on empirical transitions from a complete dataset of job changes for an entire nation. Our underlying assumption is that frequent job-to-job transitions reflect functional or contextual similarity. Unlike most existing models, our approach includes non-standard forms of employment such as part-time work, self-employment, and unemployment. This avoids the structural bias that can result from omitting transitional, informal, or flexible roles—segments that are crucial for understanding labor precarity and self-organization—and yields a more representative and temporally rich view of the labor market. Our approach also scales efficiently to large datasets, making it possible to analyze structural patterns across the full working population.

We uncover both expected and unexpected occupational links. Some communities are groupings of occupations that are functionally similar but formally classified differently. For example, one includes a range of medical tasks spanning self-employed practitioners, hospital doctors, and wholesalers of medical equipment—roles which span multiple taxonomic codes but share a professional context. Others reveal less obvious patterns, such as a community linking the full supply chain of fruit-based products: from grape and berry cultivation to beverage bottling and market sales. These examples illustrate

how empirically derived communities can reflect real-world structures that are obscured by traditional classifications.

We demonstrate the utility of these occupation communities by examining community properties, followed by two case studies. First, we examine the diffusion of educational cohorts across the labor market, capturing how different degree programs lead to narrower or broader career pathways. Second, we investigate persistent labor market disruptions following the COVID-19 lockdown, identifying occupational groups that experienced sharp exits and no full recovery in labor force share.

Our work builds on and extends several methodological traditions. Sociologically informed taxonomies such as the International Standard Classification of Occupations, 2008 revision (ISCO-08) and Nomenclature statistique des activités économiques dans la Communauté européenne (NACE), the European classification of industries (International Labour Office, 2023; Eurostat, 2008), group occupations based on expert-defined attributes like required skill and education. Mobility tables (Blau and Duncan, 1967) track flows between broad labor market classes and have been extended to capture finer-grained labor market segments (Sørensen and Grusky, 1996). More recently, network-based approaches have generalized the mobility table framework, modeling occupational connectivity based on transition probabilities, yielding insights into class structure and automation risk (Cheng and Park, 2020; Toubøl and Larsen, 2017; Rio-Chanona *et al.*, 2020; Escobari *et al.*, 2021; Villarreal, 2020).

In parallel, natural language processing (NLP) methods have been used to analyze occupational data. Some prior work translates free-text job descriptions into structured taxonomies (Kim *et al.*, 2024; Dahl *et al.*, 2024), while others use sequence embeddings to predict wages or life outcomes (Vafa *et al.*, 2024; Savcisens *et al.*, 2024). However, NLP has not yet been applied to construct empirical labor market communities from registry-based career sequences.

This paper addresses that gap. We apply a novel method that embeds occupations using word2vec trained on full career sequences, then clusters them into empirically derived labor market communities—capturing latent structural similarities from real mobility data rather than expert assumptions. We draw on the Danish Labor Market Account (Stender *et al.*, 2015), which provides monthly employment records for the entire Danish population from 2011 to

2022. These communities are then used to analyze educational cohort diffusion and COVID-19 disruption, illustrating how data-driven coarse-graining can reveal meaningful labor market structures.

3.2 Materials and Methods

We construct our empirical coarse-graining of the Danish labor market by creating embedding representations of occupations with word2vec and subsequently identifying communities through projection-based clustering. This approach allows us to capture patterns in career mobility and identify structurally similar occupations based on real-world transitions. We describe the dataset, preprocessing steps, embedding procedure, and clustering method below.

3.2.1 Data Source and Processing

Our analysis is based on the Danish Labour Market Account compiled by Statistics Denmark (Stender *et al.*, 2015), covering the period 2011–2022. The Labour Market Account provides monthly employment records for the entire Danish population—about six million individuals, of whom about 3.8 million are of working age in a given month (Statistics Denmark, 2025), including non-citizen residents. Each monthly entry includes employment status, reason for absence if relevant (e.g., sick leave, retirement), employer industry, and occupation code.

Occupations are encoded using a Danish adaptation of the ISCO-08 classification (Statistics Denmark, 2011; International Labour Office, 2023), and industries are classified using a Danish adaptation of the NACE Rev. 2 classification (Eurostat, 2008). The total dataset includes 784 occupation codes and 736 industry codes. For employed individuals, monthly working hours are recorded. When individuals hold multiple jobs simultaneously, each job is included with corresponding monthly hours worked. Demographic data such as sex, age and country of origin are also available and can be supplemented with other records using unique personal identifiers.

We restrict our analysis to individuals aged 15 to 65. Months of individual-level data where people have been registered as living outside Denmark are excluded due to inconsistent data availability. We also exclude individuals who never held a job during the study period. This leaves us with a sample of 4,373,896 individuals.

Occupational labels are based on occupation codes when available. Reporting these is mandatory for employers with at least ten employees. For employees in smaller firms (about 9% of the labor market), occupation codes are often missing; in these cases, we use the employer's industry code, which is available in 95% of instances. Self-employed individuals and co-working spouses also lack occupation codes, but are typically assigned industry codes (97% coverage). For these groups, we annotate their employment type—wage-earning, self-employed, or co-working spouse—alongside the industry label.

This labeling approach yields 2,698 distinct occupation codes based on unique combinations of occupation or industry code and employment type.

3.2.2 word2vec Representation

To capture the structure of career transitions, we train a word2vec model (Mikolov *et al.*, 2013) on sequences of occupations. Each individual's work history is transformed into a "sentence" representing their chronological occupation transitions, where "words" or "tokens" in the sentence are occupation codes. To reduce redundancy and emphasize transitions, we use only the first month of each job streak.

When individuals hold multiple jobs during the same month, we randomize job order to prevent introducing biases tied to job listing order. Extended unemployment (six months or longer) is included as a token. Similarly, periods of education are included if the individual recorded at least 10 hours of educational activity in a given month. All educational engagements are given the same token.

We compile all individual sequences into a single corpus and train a word2vec model to embed each occupation token. The model is trained for 95 epochs with a vector size of 100, context window of 7, minimum word count of 1 (to

retain rare occupations) and a downsampling rate of 0.1. These parameters were selected to minimize validation loss on a proxy prediction task; additional details are provided in word2vec Parameter Tuning.

3.2.3 Community Detection

We identify occupation communities by clustering the learned embeddings. To handle the high dimensionality of the word2vec vectors, we first project them into two dimensions using t-SNE, a technique for visualizing high-dimensional data in a lower-dimensional space that preserves local structure and is well-suited to detecting community-like groupings in the embedding space (Thrun and Ultsch, 2021).

We use t-SNE with 2 dimensions and a perplexity of 5. We then apply HDBSCAN (McInnes *et al.*, 2017), a spatial clustering algorithm, to the projected embeddings. HDBSCAN is chosen for its ability to detect communities of varying density and its robustness to noise. We use a minimum community size of 3 and a minimum sample size of 3, allowing the algorithm to identify even small but meaningful clusters. Community selection is based on the 'excess of mass' (eom) criterion.

This process yields a set of empirically derived occupation communities that serve as the foundation for our case studies. The number of communities, as well as other summary statistics, can be seen in Table 3.1.

Table 3.1.. Summary statistics of the dataset and resulting occupation communities.

Metric	Value
No. of individuals	4,373,896
No. of occupations	2,698
No. of communities	311
Unclassified occupations	355

3.3 Results

We identify 311 distinct communities. 355 occupations are not assigned to a community and thus classified as 'noise'. The community size distribution is rel-

atively uniform; the largest community contains approximately 30 occupations. A mapping of occupation codes to communities can be found in Description of Data Files.

Given the large number of communities, it is difficult to visualize their properties directly. Instead, Fig 3.1 presents a pairwise scatter matrix of key community characteristics: proportion male, mean age, age standard deviation, proportion self-employed, average salary percentile (age-adjusted), and salary percentile standard deviation (age-adjusted). Communities are colored by a four-cluster K-means clustering applied to the community property space. Point area is proportional to each community's share of the total labor market.

We observe several structured relationships between community properties. Notably, there is a negative correlation between mean age and age standard deviation: communities with older average ages tend to have narrower age distributions. This could reflect stable, long-duration careers that individuals enter later in life and remain in for longer periods. In contrast, communities with younger average ages show greater variability, which may indicate transitional occupations, broader entry pathways, or less standardized career timing.

Gender composition also shows an association with average income levels. Communities with the highest average salary percentiles tend to be male-dominated, while female-dominated communities are largely absent from the upper end of the income distribution. Since salaries are age-adjusted (based on percentiles within age groups), this pattern cannot be explained by differences in experience alone. Instead, it suggests systematic gender-based differences in occupational sorting or compensation structures.

Another prominent pattern involves the standard deviation of salary percentiles within communities. Some communities exhibit high internal wage dispersion even after adjusting for age, suggesting that they contain a wide range of roles or pay structures within the same occupational cluster. This is especially true for communities with moderate levels of self-employment, where the employed subgroup may span both high- and low-paying roles, or where conventional employment mirrors a more heterogeneous occupational landscape. In contrast, many large communities with predominantly wage-earning workers

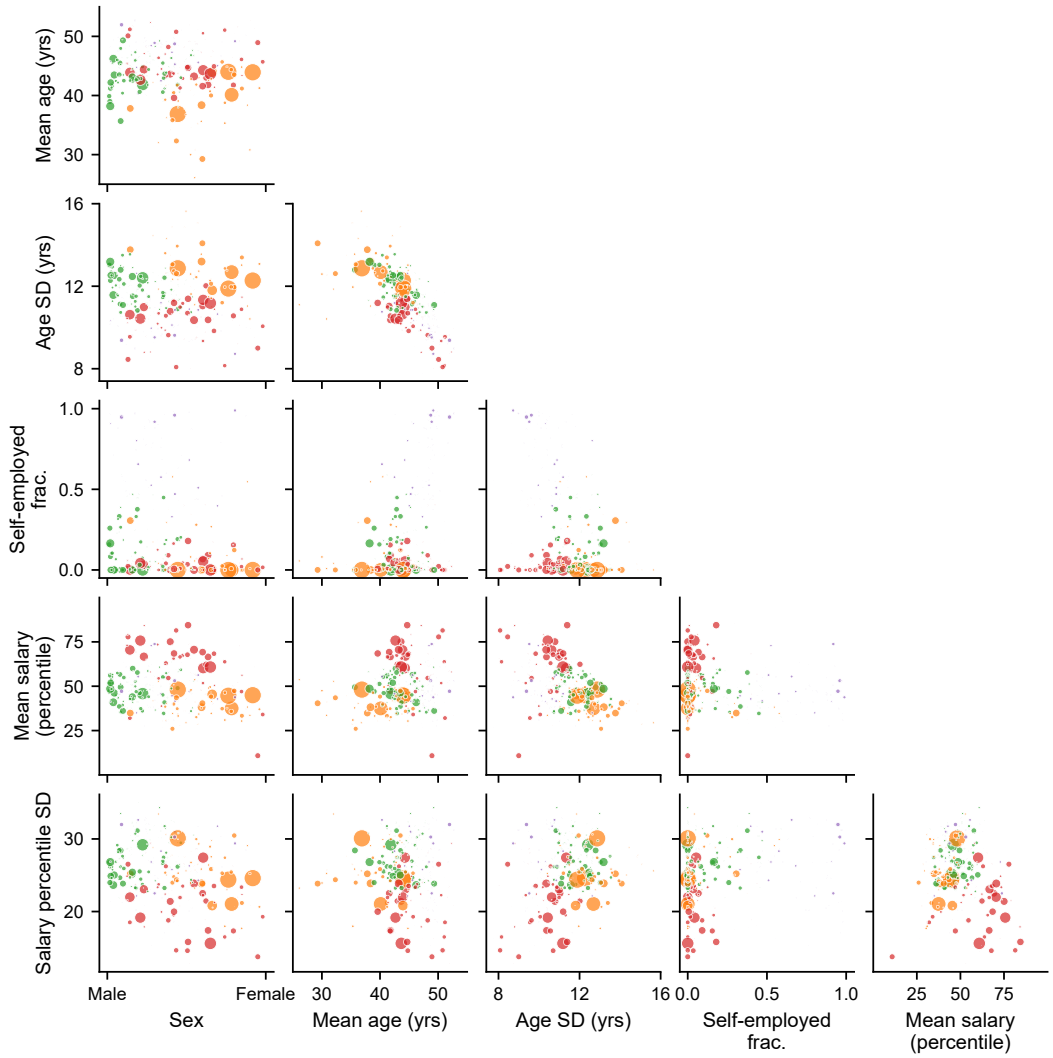

Figure 3.1.. Pairwise scatter matrix of community properties. Proportion male/female, mean age, standard deviation of age, proportion self-employed, average salary percentile (by age brackets) and salary percentile standard deviation (by age bracket). Colors represent the K-means clustering of the community property space using four means. Point size represents the proportion of total labor market hours in that community.

exhibit tighter salary distributions, suggesting more uniform compensation structures.

Having defined the communities, we now go through three concrete use-cases to demonstrate their usefulness for understanding the labor market.

3.3.1 Community Properties

To interpret the structure of the derived occupation communities, we examine their demographic and economic characteristics over time. This allows us to understand who inhabits each community and how these groups evolve.

As an example, in Fig 3.2 we have plotted community properties of three representative communities, each of which had at least 1% labor market share at some point in the study period: legal work, early childhood education, and animal farming. These three communities were selected to represent different labor sectors: office work, care work, and agriculture. Description of Data Files contains corresponding data for all communities.

For each community, we examine three properties: the share of work hours in that community relative to the labor market as a whole, the share of labor hours in that community done by men compared to the labor market as a whole, and the share of labor in the community done as part of self-employment or working in a spouses' business. Each time series has been seasonally de-trended.

Next, we explore the age and salary composition of a community over time, focusing on the legal work community as an example (for data on other communities, please refer to Description of Data Files, Description of Data Files, Description of Data Files, and Description of Data Files). Fig 3.3A displays the normalized age distribution for this community expressed relative to the labor market as a whole. White-collar communities such as this one often show cohort-based entry patterns, with distinct bands of workers entering and exiting at similar ages rather than a steady inflow. They also tend to exhibit later retirement than more physically demanding occupations (Brydsten *et al.*, 2023; Undem *et al.*, 2025). This may explain the overrepresentation of older legal workers in 2012. By 2022, this pattern fades—not necessarily because

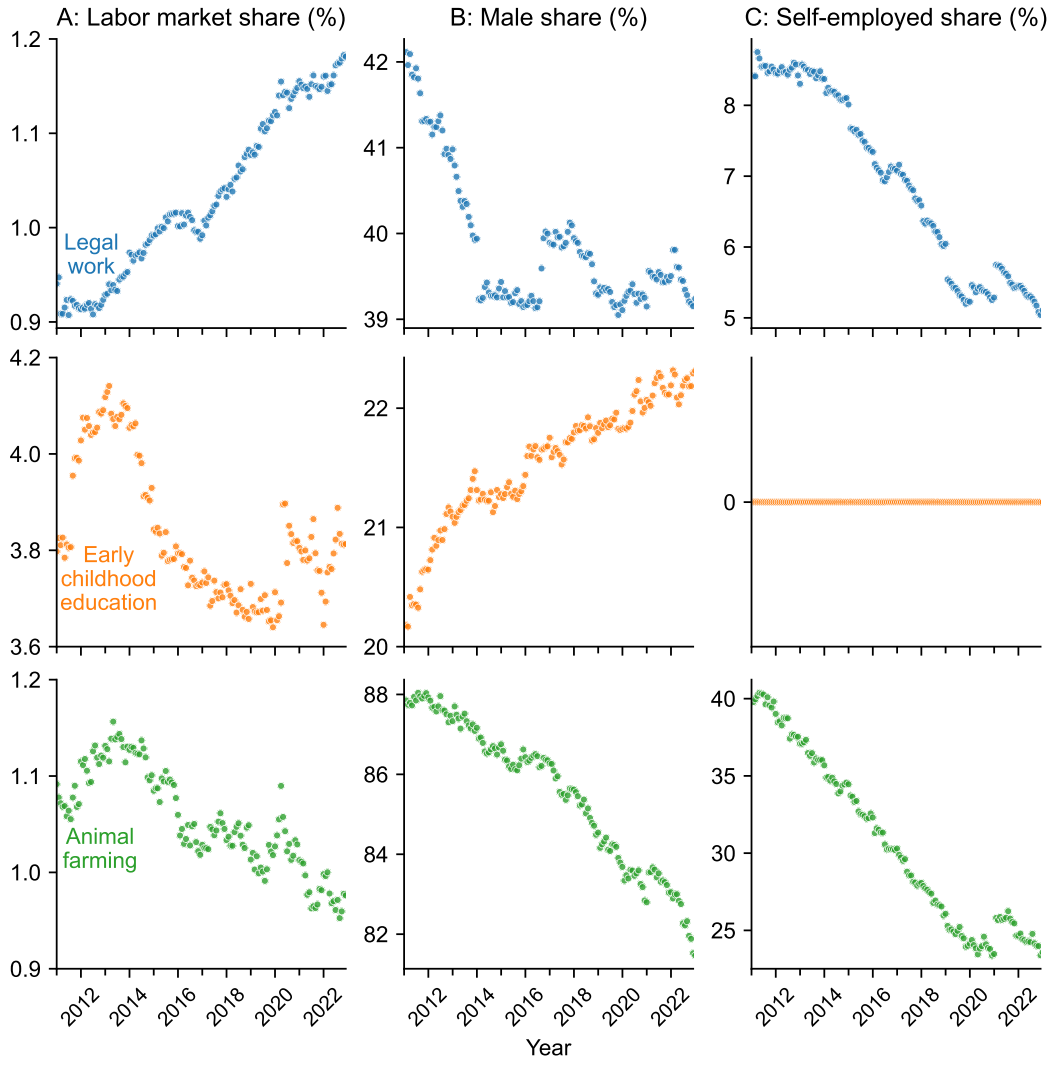

Figure 3.2.. Demographic and employment properties of three communities. (A) Labor market share (the total labor market hours worked in the community as a percentage of labor market hours as a whole). (B) Male share (the percentage of labor hours worked in the community by men). (C) Self-employed share (the percentage of labor hours worked in the community by self-employed individuals and co-working spouses). Each variable is represented for three different communities (legal work, early childhood education and animal farming) over time. All subplots have been seasonally de-trended. Groups with fewer than three individuals in either the group or its complement were excluded to protect privacy; flat zeros may therefore indicate suppression rather than literal absence.

legal workers retired earlier, but perhaps because physically demanding jobs had become less common in later cohorts.

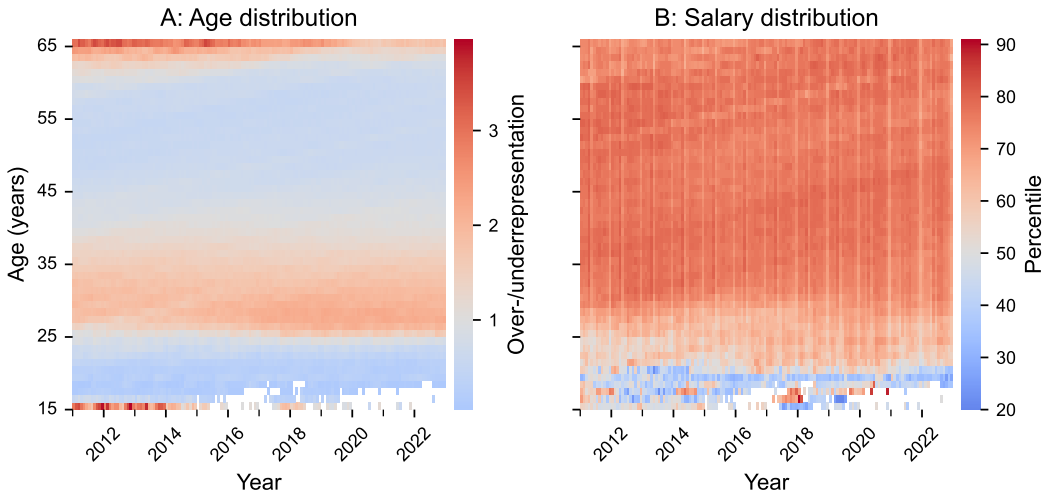

Figure 3.3.. Age and salary distribution in legal work over time. (A) Age over-/underrepresentation in legal work over time. A value of 1 indicates representation on par with the full labor market, a value of 2 indicates representation double that of the full labor market, and so on. (B) Salary distribution in legal work by age over time, expressed in terms of percentile rank of individuals of the same age at the same time in the labor market as a whole. To protect anonymity, cells were removed from either subplot if they represented the values of fewer than 3 individuals.

Earnings provide another important lens for understanding occupational communities. For wage earners (excluding the self-employed and benefit recipients), we compute monthly hourly wages and assign each person a percentile rank within their age group for that month. Fig 3.3B shows the resulting distribution of salary percentiles over time for the legal work occupation community, which indicates that individuals in this community are often higher-earning than average.

3.3.2 Diffusion

Using the communities, we can study how individuals with a specific educational background spread throughout the labor market. We refer to this process as 'diffusion'. By tracking groups of individuals who completed the same degree under comparable conditions, we can assess the heterogeneity of their labor market outcomes over time.

We examine nine educational cohorts: five professional degrees (Smithing, Finance, Computer Science, Film and TV Production, Nursing) and four longer academic programs (Business Economics, Computer Science, Danish, and Medicine). Note that in the Danish context, professional and academic Computer Science programs differ substantially: the former is a shorter, practice-oriented vocational degree, while the latter is a university degree with a stronger theoretical emphasis. As a baseline, we also include a random sample of 2% of individuals who completed any higher education in the same year (1309 people).

For each degree cohort, we include only those who completed the program in 2010—the year before our observation window begins—and restrict the sample to individuals whose time to degree completion falls within one standard deviation of their cohort mean. For the random sample, we remove broad outliers, but do not apply degree-specific filters due to varying program lengths.

To quantify diffusion, we calculate *temporal entropy*, a measure of how dispersed each cohort becomes across occupation communities over time. At each time point t, we compute the entropy of the community distribution:

$$H_t = -\sum_{i=1}^{k} p_{i,t} \log(p_{i,t}),$$

where $p_{i,t}$ is the fraction of the cohort in community i at time t. Higher entropy indicates greater diversity in labor market outcomes.

Fig 3.4A shows strong differences in how educational cohorts spread across the labor market over time. (A version of this figure without aggregation over communities is included in Temporal Entropy, Disaggregated View for comparison.) The entropy scale (measured in bits) reflects the effective number of communities occupied by each cohort at each time point; for example, an entropy of 3 corresponds to a spread across $2^3 = 8$ communities. The randomly sampled baseline cohort reaches above 5 bits early in the period, reflecting wide and sustained occupational diversity. In contrast, graduates from structured programs like medicine and nursing remain highly concentrated, with entropy staying below 1.5 bits even after a decade. This suggests that their labor market pathways are narrow and stable. Other programs—such as

smithing, film and TV production, and computer science—exhibit moderate and gradually increasing entropy, reflecting more varied outcomes. Business economics stands out for its continued rise in entropy throughout the observation period, indicating not only broad dispersion but increasing heterogeneity over time.

To assess consistency in diffusion patterns, we repeat the entropy analysis across multiple graduation years for the Nursing cohort, which leads to one of the most common single occupations in Denmark (approximately 2% of the workforce, excluding the self-employed and small business employees). Fig 3.4B shows the results. Across cohorts, we observe a consistent pattern: high and seasonally varying entropy during the education period (capturing student job variability), followed by a sharp drop at labor market entry, and then a gradual increase in entropy as careers evolve. Notably, this structure is only visible because we retain student job data in our sequences.

3.3.3 COVID-19 Effects

In our final case study we analyze the effects of COVID-19 on the labor market. The pandemic caused a large disruption to labor markets, but it remains unclear to what extent this disruption was temporary or resulted in lasting structural change. By tracking the proportion of individuals in each community over time, we can identify which occupational communities were most affected by the COVID-19 pandemic. Many communities show a visible decline in labor market share during the early months of lockdown, followed in most cases by a recovery toward pre-pandemic levels. However, a small number of communities show a persistent drop with no full return to previous trends, suggesting that some types of work, and the people who perform them, may be more vulnerable to long-term displacement.

In particular, we find that only three communities show no recovery to pre-COVID levels: drivers of cars (e.g., taxi drivers and chauffeurs), bus drivers, and airport workers. Leveraging the full labor market histories in our dataset, we examine churn within these communities, defined as individuals entering or exiting the community. For drivers, Fig 3.5A shows a pronounced spike in exits from the community to outside the labor force at the onset of the Danish lockdown in March 2020. Notably, this spike is not matched by a

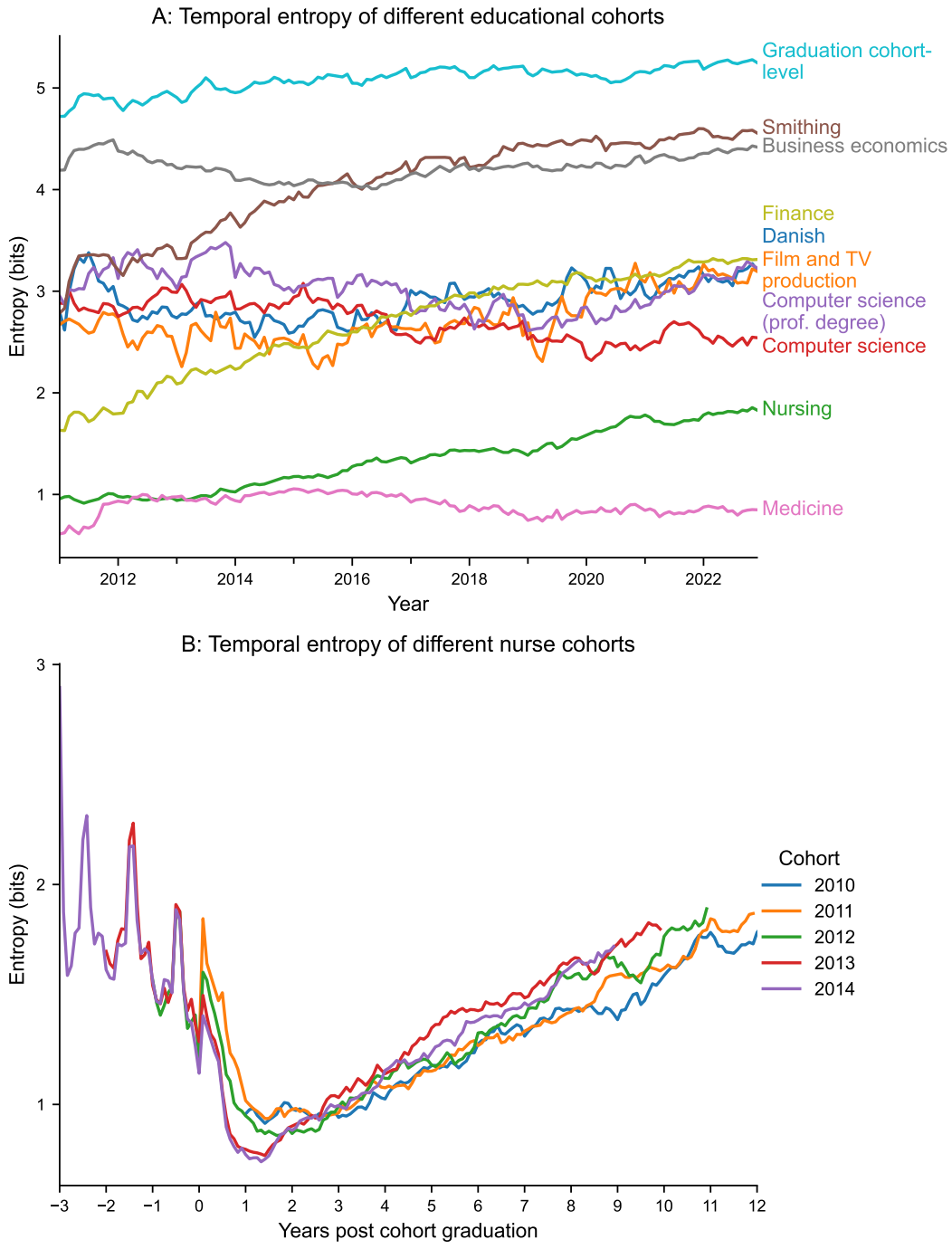

Figure 3.4.. Temporal entropy of different educational cohorts. (A) Temporal entropy for nine different educational cohorts who completed a degree in the indicated subject in 2010, plus a random sample of 2% of graduates in that year. (B) Temporal entropy for individuals who completed a nursing education in years 2010–2014 as a function of years post education.

comparable wave of re-entry, even as restrictions ease. There is also some evidence of increased movement from this community into other communities after 2020, though the available follow-up period is too short to confirm a long-term trend.

To further quantify churn, we compute the Jaccard index between consecutive months: the ratio of individuals remaining in the community from one month to the next. As shown in Fig 3.5B, the driver community exhibits a stable pattern prior to the pandemic, followed by a sharp decline at the time of lockdown. While some recovery occurs, the index remains lower than pre-pandemic levels, suggesting sustained changes in workforce stability or occupational structure. Whether this reflects a permanent shift in employment patterns for drivers remains to be seen.

3.4 Discussion

In this study, we introduced an empirical coarse-graining of the Danish labor market using word2vec embeddings of career sequences, followed by projection-based clustering. This approach allowed us to group occupations into meaningful communities based on observed mobility patterns. We demonstrated the utility of these communities through several applications: profiling the demographic and economic characteristics of communities, analyzing diffusion of educational cohorts into the labor market, and assessing the long-term effects of the COVID-19 pandemic on specific occupational groups.

Our approach has several strengths. First, it captures job similarity from observed transitions rather than expert-defined taxonomies, making it sensitive to real-world labor dynamics. Second, it flexibly incorporates part-time, student, and self-employment states, as well as worker histories, which are often excluded or handled inconsistently in other studies. Third, the method scales well to the size of the registry-based Danish dataset, which provides population-level monthly granularity, enabling fine-grained temporal analyses not feasible in many countries.

That said, some limitations remain. The time window (2011–2022) covers a relatively short period, and our model aggregates transition behavior across

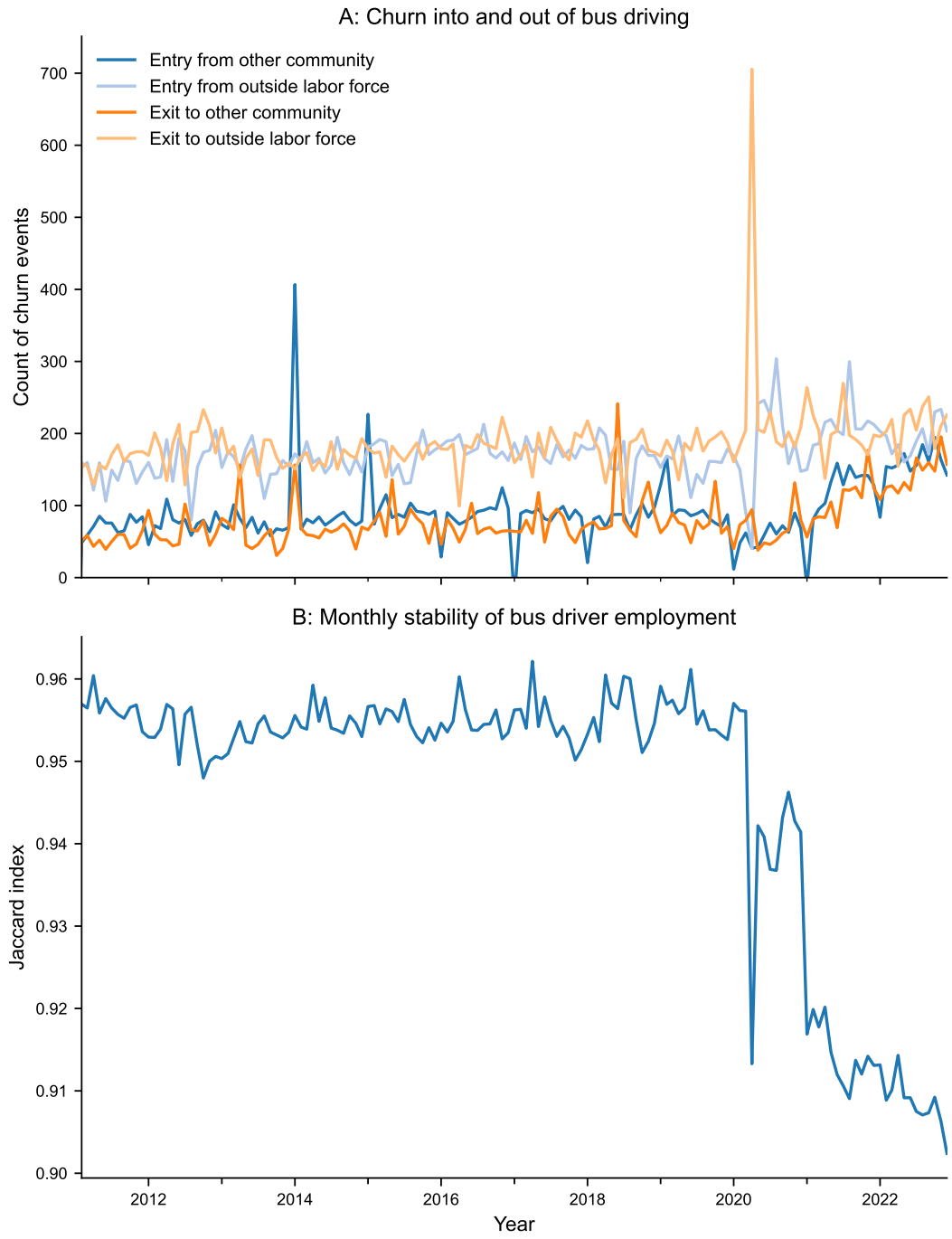

Figure 3.5.. Churn of drivers over time. (A) Churn in the occupation community of drivers as a function of year (seasonally detrended), decomposed into churn into and out of the labor force, and into and out of a different communities. (B) Jaccard index of the set of people in this community over consecutive months.

it. This likely obscures temporal shifts in labor mobility, especially during periods of major disruption such as the COVID-19 pandemic. A natural next step would be to explore dynamic coarse-graining methods (e.g., temporal community detection or latent space tracking as seen in Dabke and Dorabiala, 2024; Karaaslanli and Aviyente, 2019; Sewell and Chen, 2017) to account for evolving occupational structures.

Additionally, the dataset excludes informal labor, which accounts for an estimated 9% of private sector labor input in Denmark as of 2019. Although most of this informal work consists of small-scale one-off tasks (ELA, 2023), some long-term employment relationships may go unrecorded. Finally, the data contain inconsistencies or gaps in occupation labeling and income data, particularly for self-employed individuals and co-working spouses, whose job roles and hours are hard to measure.

Despite these limitations, the results demonstrate that data-driven occupation coarse-graining can reveal interpretable labor market structures and support a wide range of empirical investigations. This approach offers a flexible and scalable foundation for future work on labor market dynamics, economic shocks, and career outcomes.

References

Blau, Peter M. and Otis Dudley Duncan (1967). *The American Occupational Structure*. New York: John Wiley and Sons.

Brydsten, Anna, Caroline Hasselgren, Mikael Stattin, and Daniel Larsson (Sept. 2023). „The Road to Retirement: A Life Course Perspective on Labor Market Trajectories and Retirement Behaviors“. In: *Work, Aging and Retirement* 11.1, pp. 1–12.

Cahuc, Pierre, Stéphane Carcillo, and André Zylberberg (2014). „Equilibrium Unemployment“. In: *Labor Economics*. 2nd. The MIT Press, pp. 553–574.

Cheng, Siwei and Barum Park (Nov. 2020). „Flows and Boundaries: A Network Approach to Studying Occupational Mobility in the Labor Market“. In: *American Journal of Sociology* 126, pp. 577–631.

Dabke, Devavrat Vivek and Olga Dorabiala (Dec. 2024). „Vertex clustering in diverse dynamic networks“. In: *PLOS Complex Systems* 1.4, pp. 1–29.

Dahl, Christian Møller, Torben Johansen, and Christian Vedel (2024). *Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE*. arXiv: 2402.13604.

ELA (Mar. 2023). *Factsheet on undeclared work – DENMARK*. Tech. rep. European Labour Authority.

Escobari, Marcela, Ian Seyal, and Carlos Baboin (June 2021). *Moving up: Promoting workers' upward mobility using network analysis*. Tech. rep. Brookings.

Eurostat (2008). *NACE rev. 2: Statistical classification of economic activities in the European Community*. Luxembourg: European Commission.

International Labour Office (2023). *The International Standard Classification of Occupations (ISCO-08) Companion Guide*. Geneva: International Labour Office.

Kalleberg, Arne L. (2009). „Precarious Work, Insecure Workers: Employment Relations in Transition". In: *American Sociological Review* 74.1, pp. 1–22.

Karaaslanli, Abdullah and Selin Aviyente (2019). *Constrained Spectral Clustering for Dynamic Community Detection*. arXiv: 1911.01475.

Kim, Tae-Yeon, Seong-Uk Baek, Myeong-Hun Lim, *et al.* (2024). „Occupation classification model based on DistilKoBERT: using the 5th and 6th Korean Working Condition Surveys". In: *Annals of Occupational and Environmental Medicine* 36, e19.

McInnes, Leland, John Healy, and Steve Astels (Mar. 2017). „hdbscan: Hierarchical density based clustering". In: *Journal of Open Source Software* 2.11, p. 205.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781.

Rio-Chanona, R. Maria del, Penny Mealy, Mariano Beguerisse-Díaz, François Lafond, and J. Doyne Farmer (2020). „Occupational mobility and automation: a data-driven network model". In: *Journal of the Royal Society Interface*.

Savcisens, Germans, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann (2024). „Using sequences of life-events to predict human lives". In: *Nature Computational Science* 4.1, pp. 43–56.

Sewell, Daniel K. and Yuguo Chen (June 2017). „Latent Space Approaches to Community Detection in Dynamic Networks". In: *Bayesian Analysis* 12.2.

Sørensen, Jesper B. and David B. Grusky (1996). „The Structure of Career Mobility in Microscopic Perspective". In: *Social Differentiation And Social Inequality*. Routledge.

Statistics Denmark (Mar. 2011). *DISCO-08: Danmarks Statistiks fagklassifikation*. Technical manual. First edition. Statistics Denmark.

Statistics Denmark (2025). *Population Figures*. https://www.dst.dk/en/Statistik/emner/borgere/befolkning/befolkningstal. Accessed: 2025-05-05.

Stender, Pernille, Thomas Thorsen, and Hans Henrik Andersen (2015). „Micro data integration for Labour Market Account". In: *Statistical Journal of the IAOS*.

Thrun, M. C. and A. Ultsch (2021). „Using Projection-Based Clustering to Find Distance- and Density-Based Clusters in High-Dimensional Data". In: *J Classif* 38, pp. 280–312.

Toubøl, Jonas and Anton Grau Larsen (2017). „Mapping the Social Class Structure: From Occupational Mobility to Social Class Categories Using Network Analysis". In: *Sociology* 51.6, pp. 1257–1276.

Undem, Karina, Taina Leinonen, Daniel Falkstedt, Gun Johansson, Jacob Pedersen, Eira Viikari-Juntura, Ingrid Sivesind Mehlum, and Svetlana Solovieva (July 2025). „Occupational differences in working life expectancy and working years lost in Nordic countries". In: *Scandinavian Journal of Work, Environment & Health*.

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei (2024). *CAREER: A Foundation Model for Labor Sequence Data*. arXiv: 2202.08370.

Villarreal, Andrés (2020). „The U.S. Occupational Structure: A Social Network Approach". In: *Sociological Science* 7.8, pp. 187–221.

4

Occupations in Contemporary Popular Fiction

This chapter is based on the manuscript:
Ella Clement.
Imagining Work: Professions, Prestige, and Desire in Bestselling Fiction. Submitted to Poetics, 2025.
A preprint of this manuscript is available on SocArXiv: `https://osf.io/preprints/socarxiv/qaks4_v1?view_only=`, with DOI: `https://doi.org/10.31235/osf.io/qaks4_v1`.

Abstract

I analyze how occupations are portrayed in contemporary popular fiction by coding the protagonists' and love interests' professions in 1181 bestselling novels from 2023–2024. Using large language model–assisted extraction with human oversight, I map which roles dominate across genres and compare them to real-world occupational data. The results show that fiction highlights a narrow set of archetypes—e.g., students, writers, detectives, athletes—while downplaying most everyday forms of labor. Overrepresentation is shaped partly by prestige: higher-status jobs are somewhat more likely to be emphasized, but many exceptions remain, with some prestigious professions underrepresented (e.g., tech executives and engineers) while low-prestige but narratively useful roles (e.g., hunters, antique dealers) are prominent. In romance titles, occupational portrayals diverge sharply by gender: men are written into positions of power, danger, or transgression, while women are more often portrayed in expressive or domestic roles. Comparing these fictional portrayals to survey and dating-app data on occupational desirability reveals consistent gaps, underscoring how cultural products use work as sym-

bolic resources for storytelling and signaling desirability rather than mirroring labor markets.

4.1 Introduction

Fiction is not only a source of entertainment but also a cultural space where the meanings of work, status, and desirability are constructed and circulated (Couldry, 2012). The professions assigned to protagonists and love interests in novels thus signal more than just character backstory: they reveal symbolic hierarchies of labor, prestige, and aspiration. A detective, a doctor, or a florist is never a neutral occupational choice but a shorthand for competence, authority, glamour, or ordinariness.

Sociological research has long been concerned with the prestige of occupations (Treiman, 1977). Surveys rank jobs by public esteem, linking occupational status to education and earnings (Hughes *et al.*, 2024; Valentino, 2019). At the same time, cultural sociologists have examined how symbolic boundaries such as gender and class shape portrayals of identity in media (Holtzman and Sharpe, 2014). Studies of romance novels, for example, show how genre conventions reproduce and reshape ideals of gender and desire (Radway, 1991). Research on assortative mating and dating preferences demonstrates that social status and earnings are key dimensions of attractiveness (Buss, 1989; Shackelford *et al.*, 2005). Finally, scholars of work have emphasized the invisibility of many forms of everyday labor in cultural representation, especially service and care work (Glenn, 1992; England, 2005).

Yet despite these strands, we lack systematic evidence on how professions are distributed across contemporary fiction. Prior analyses of media occupations have often been qualitative (e.g., Zoonen, 1996), restricted to particular genres (e.g., Regis, 2003), or focused on film and television rather than novels (e.g., Coltrane and Adams, 1997). Quantitative studies of occupational prestige, meanwhile, capture attitudes toward real-world work but do not extend to how such roles are used as symbolic shorthands in cultural products. Nor has research on occupational desirability in romantic partners been connected to the fictional portrayals that circulate most widely in popular culture.

This article addresses these gaps through a large-scale analysis of professions in 1181 novels that appeared on Anglo-American bestseller lists during 2023–2024. I focus on commercially successful titles because they capture occupational portrayals most visible in mass culture, rather than the narrower worlds of literary prestige or experimental fiction. Using large language model–assisted coding combined with manual validation, I identify protagonists' and (where relevant) love interests' professions across genres. I then compare fictional prevalence to real-world occupational statistics, examine the relationship between representation and occupational prestige, and analyze how romance and fantasy romance in particular encode gendered desirability scripts.

The findings reveal patterned distortions: fiction privileges rare or symbolically charged occupations while underrepresenting the everyday work that sustains most lives. Prestige plays a role but is not determinative: high-status jobs such as tech executives or engineers are often sidelined, while lower-prestige but narratively useful roles such as hunters or crime bosses are foregrounded. Romance genres further highlight gender asymmetries, with men portrayed in roles tied to power, danger, and transgression, while women appear in roles emphasizing youth, creativity, or domesticity. These divergences between fiction, real-world prevalence, and stated romantic partner preferences highlight how occupations function as a symbolic resource in storytelling, reflecting cultural narratives of work, status, and desire.

4.2 Materials and Methods

4.2.1 Fictional Prevalence Dataset

I began by compiling a list of all bestselling fiction titles from 2023—2024 by scraping the adult/general fiction bestseller lists on Publishers Marketplace.[1] I supplemented these data with book descriptions and genre classifications from the Google Books API. These classifications were used only to distinguish novels from other works (e.g., nonfiction, poetry, or graphic novels). When the

[1]The lists I used were ABA Indiebound's Hardcover Fiction and Paperback Fiction, Amazon Combined Print/eBook Fiction, and the New York Times Hardcover Fiction and Paperback Fiction.

Google Books label was ambiguous or inconsistent with the book description, I checked Goodreads categories to resolve the discrepancy.

The unit of analysis throughout is the unique novel title. Duplicate entries across bestseller lists and multiple editions of the same work were consolidated. I excluded any titles that were not novels—such as short story collections, graphic novels, and poetry volumes (there were 180 removed titles; see linked GitHub repository for a full list). A small number of non-novel works that escaped initial filtering were removed during validation. After this cleaning process, the dataset comprised 1181 English-language bestselling novels. Where available, I enriched descriptions with full plot summaries scraped from SuperSummary (available for 533 titles). The full dataset, as well as the code to generate it, are available from `https://github.com/ella-clement/fiction-professions`.

To extract structured information—including genre, protagonist(s), love interest(s) where applicable, and their respective professions—I queried GPT-4o via the OpenAI API (OpenAI, 2024) using a multi-stage prompting strategy that combined model knowledge with external summaries (see B.1 for details). In cases with multiple protagonists or an ensemble cast, I coded all central figures. When no occupation was specified, the field was coded as "None." I also introduced categories for common cases in which traditional employment was irrelevant (e.g., 'child,' 'student,' 'retired').

Narrower genre classifications within the novel corpus (e.g., romance, fantasy, thriller, literary fiction) were generated through GPT-4o and then standardized manually. This process included unifying synonymous labels (e.g., "fantasy romance" and "romantasy") and folding specific tags (e.g., "dystopian") into broader categories (e.g., science fiction). I manually incorporated additional input from Goodreads.

For the subset of romance and fantasy romance novels, I additionally hand-coded the gender of protagonists and love interests as men or women, based on blurbs and, where necessary, the book itself; here "gender" refers to characters' gender as presented in the texts. This binary coding reflects the fact that these genres overwhelmingly portray characters in binary gender terms, and gendered portrayals are central to the analysis of desirability.

To evaluate the accuracy of automated coding, I drew a random 10% subsample of novels ($n = 117$). Half of this subsample was used to refine coding strategies (test set), and the remaining half was reserved for independent accuracy checks (validation set). For each sampled book, I read the blurb or, when necessary, the book itself to identify genre, protagonist name(s) and profession(s), and, for romance and fantasy romance titles, love interest name(s) and profession(s). This combination of automated extraction with iterative human validation follows the hybrid logic of computational grounded theory (Nelson, 2020).

Accuracy metrics are presented in Table 4.1. The automated coding achieved 92–95% accuracy for protagonist identification and 80–83% for professions, levels comparable to intercoder reliability benchmarks in manual content analysis (typically around .80 or above; see Neuendorf, 2017). Accuracy was generally higher for well-known books, particularly those with detailed summaries and straightforward narrative structures involving a single protagonist with a clearly defined occupation. The most common inaccuracy was for the model to return that information was unknown; this error accounted for about 77% of all inaccuracies in profession.

Table 4.1.. Performance metrics on validation and test sets.

Dataset	No. Books[a]	Genre Accuracy (%)	Protagonist Identified (%)	Protagonist Profession (%)	Love Interest Identified[b] (%)	Love Interest Profession[b] (%)
Test Set	57	100	94.7	79.8	100	93.3
Validation Set	60	100	92.5	82.5	100	71.4

[a]Book counts differ because several test set books were removed post hoc after being deemed ineligible for the sample.
[b]Love interest identification and profession accuracy are based only on books in the romance and fantasy romance genres. This was 15 books in the test and 21 books in the validation sets, respectively.

For the final dataset used in analysis, I incorporated the hand-coded information from the validation sample back into the corpus, replacing the corresponding automated outputs. This ensured that the 10% of novels I reviewed directly reflected the more accurate human coding, while the remainder of the dataset retained the automated results.

4.2.2 External Comparison Datasets

To situate fictional occupations against real-world benchmarks, I drew on three types of external data: U.S. occupational prevalence and wages, survey-based prestige scores, and measures of occupational desirability in romantic partners. Each dataset required harmonization with the fictional data and carried specific limitations, as detailed below.

Real-world prevalence estimates

In Section 4.3.2, I compare fictional representation to U.S. employment statistics. The primary source was (U.S. Bureau of Labor Statistics, 2024), which reports national counts and average hourly wages by occupation. To align my dataset to this source, I aggregated or collapsed categories where necessary (e.g., combining ”novelist” and ”poet” under ”writer”), and in some cases merged real-world categories upward (e.g., grouping all postsecondary teachers when the subject field could not be distinguished in fiction).

Not all fictional professions mapped onto real-world categories. Some were non-employment states (e.g., student, retiree), and others were vanishingly rare or illegal (e.g., bounty hunter, crime boss). Where possible, I used secondary online sources to obtain approximate U.S. counts; in several cases these should be understood as order-of-magnitude estimates only, which is acceptable given that fictional and real-world prevalence are compared on a log scale. For a table of these secondary sources, please see Table B.1.

Professions for which no plausible estimate could be found were omitted from analysis, as were real-world occupations that never appeared in fiction. For wages, I took weighted averages across combined categories and supplemented missing values with estimates from Zippia (*Zippia* 2025) or Glassdoor (*Glassdoor* 2025).

This dataset provides a baseline against which to judge whether fictional occupations reflect labor market frequency or instead highlight symbolically salient roles.

Prestige data

In Section 4.3.3, I compare fictional representation to survey-based occupational prestige scores from (Hughes *et al.*, 2024). This survey was conducted via MTurk in 2015–2016, and involved 3076 respondents who rated the prestige of 1029 occupations on a scale from 0 to 100. While MTurk samples are not nationally representative, they are widely used in social science research (Berinsky *et al.*, 2012), and the resulting scores align with long-established prestige scales (Treiman, 1977). Because this comparison requires both a real-world prevalence estimate and a prestige score, the resulting analysis is limited to the subset of occupations available in both datasets.

Romantic partner preference data

In Section 4.3.4, I compare occupational portrayals in romance novels to three datasets reporting real-world rankings of desirability in romantic partners by gender and occupation. Two of these datasets come from Tinder samples that report match rates for profiles listing different professions (Binder, 2016; Education Connection, 2022), while the third is a Zippia survey of career desirability (Morris, 2021). These sources vary in coverage: the Tinder samples reflect only users who disclose occupations and are shaped by the demographics of the app, while the Zippia survey is limited to 250 common jobs and lacks details on sample composition. They are also all limited to preferences in heterosexual pairings. While none of these sources constitute a representative academic survey, together they provide a pragmatic proxy for occupational attractiveness.

Because the datasets list different sets of occupations, rank correlation tests would be uninformative. Instead, I focus on the six top-ranked professions in each, which aligns with the shortest of the lists, in this case (Education Connection, 2022). I did not attempt to harmonize occupational labels across sources, as the full set of occupations under consideration in the three real-world sources is unavailable.

Taken together, these three external data sources—on labor market prevalence, prestige, and occupational desirability in romantic partners—provide complementary benchmarks. They allow fictional portrayals to be situated not just against frequency, but also against status and cultural ideals of attractiveness.

4.3 Results

4.3.1 Overall Prevalence

Figure 4.1 shows the thirty most common professions across all novels. The distribution is top-heavy: a handful of roles dominate disproportionately, especially students, writers, and detectives. Many of these are not everyday jobs but narrative archetypes (e.g., detective, secret agent, royalty, warrior, assassin), reflecting their utility as plot engines. Cultural and creative roles such as writers, musicians, and artists are also highly visible, signaling the importance of creativity and self-expression in fiction. By contrast, authority professions like doctors, lawyers, scientists, and teachers appear regularly as figures of institutional credibility. Equally striking are the absences: common occupations in retail, service, or technical fields do not appear in the top thirty.

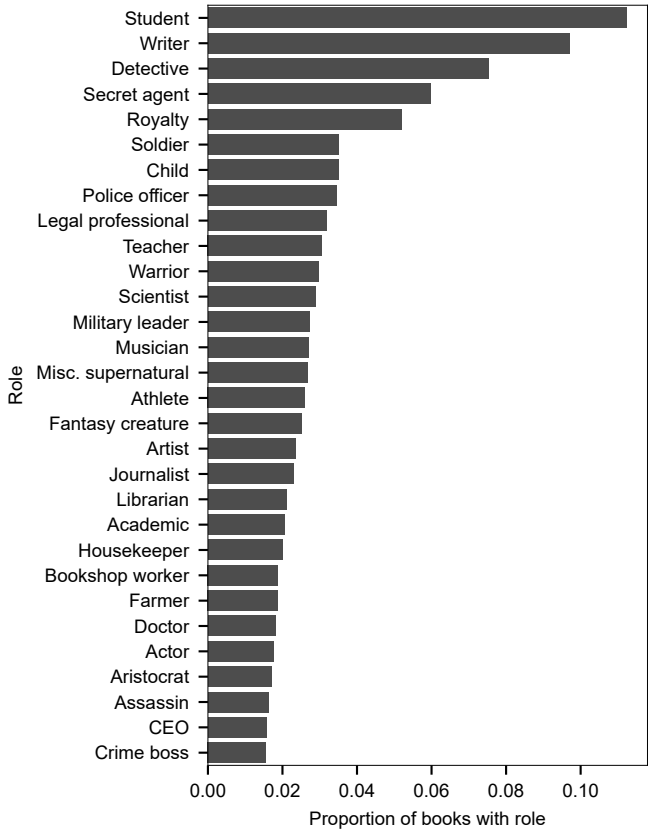

Figure 4.1.. The thirty most common roles in popular novels. Roles include both occupations and non-employment statuses.

Figure 4.2 breaks these occupations down by genre, revealing how some roles are tightly bound to genre conventions while others cut across boundaries. Detectives anchor mystery, secret agents cluster in thrillers, royalty in fantasy romance, and warriors in fantasy. These localized roles function as genre-defining cues (Cawelti, 1976), while more flexible occupations such as students and writers appear broadly across genres. Together, the figures suggest that fiction does not distribute labor evenly but sorts professions into categories: some genre-defining, others which span genres, and many as invisible background work.

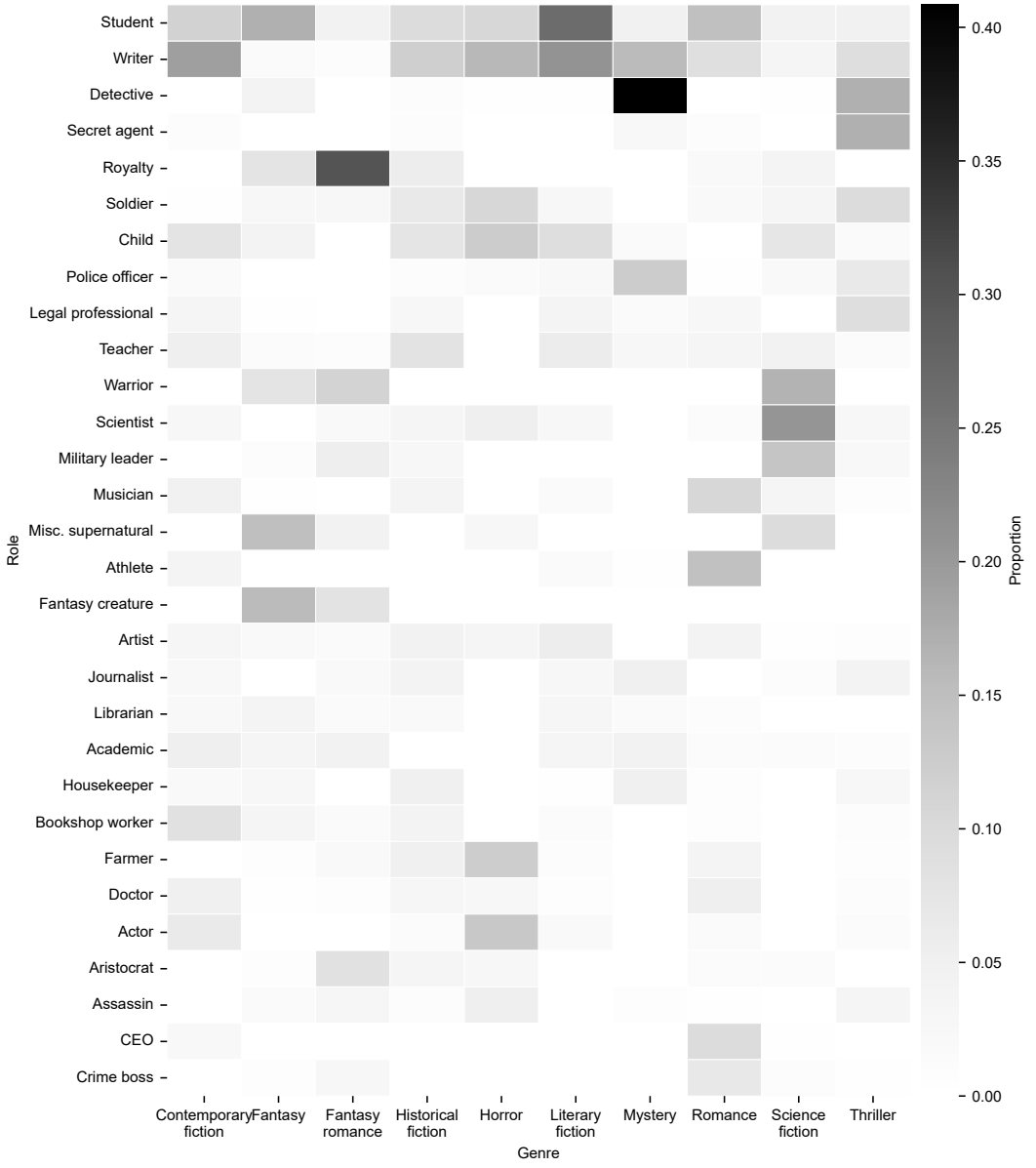

Figure 4.2.. The thirty most common roles in popular novels by genre. Roles include both occupations and non-employment statuses.

These descriptive counts suggest that fiction highlights occupations for their narrative weight rather than their real-world prevalence. To assess this, I directly compare fictional and real-world data in the next subsection.

4.3.2 Comparison to Real-world Prevalence

Figure 4.3 shows fictional representation versus real-world prevalence. Some occupations fall near parity with real prevalence, but the departures are telling. The most overrepresented roles cluster at symbolic extremes—figures of power or exception (royalty, aristocrat, astronaut), danger (assassin, bounty hunter), or glamour (athlete, secret agent). Conversely, the most underrepresented are forms of ordinary or 'invisible' labor, such as administrators, gig workers, or handypersons.

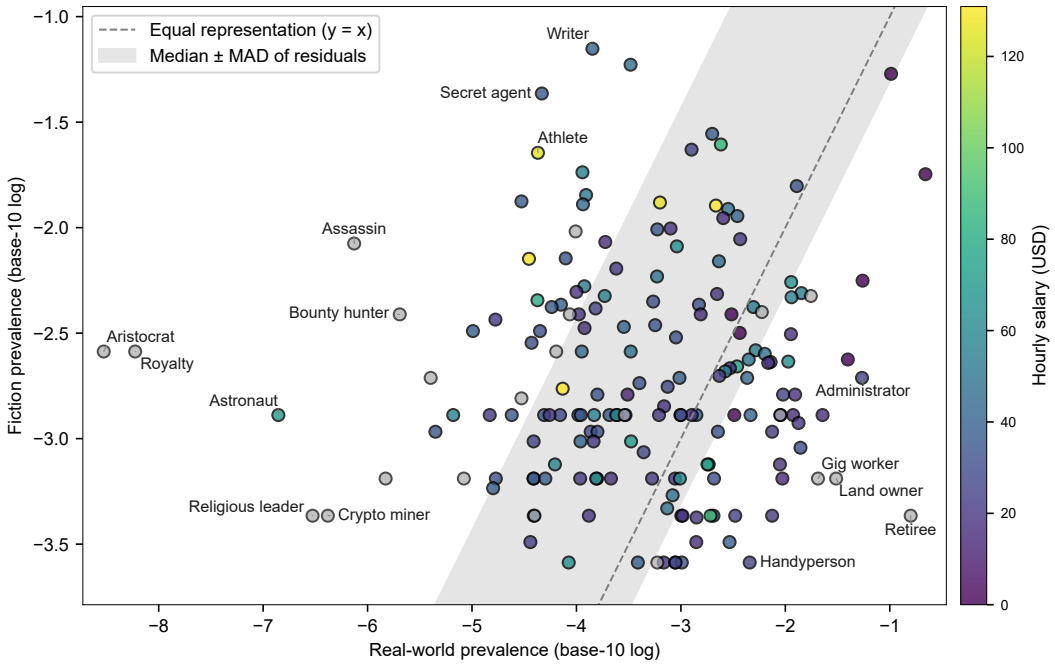

Figure 4.3.. Fictional prevalence vs. real-world prevalence. Prevalence of roles in genres set in a contemporary real-world context (contemporary fiction, horror, literary fiction, mystery, romance, and thrillers) versus U.S. employment statistics. Points are colored by hourly salary where available; gray indicates missing salary data. The five most overrepresented and underrepresented roles are labeled. Both axes use base-10 log scale.

Points are colored by mean hourly wage. Higher-paying jobs show a modest tendency toward overrepresentation (Spearman's $\rho = 0.30$, $p < 0.001$), but narrative salience is not reducible to earnings: several well-paid but unglamorous

roles fall below parity, while symbolically charged occupations with average or low wages are elevated.

Prevalence alone, however, may not fully explain why certain roles capture disproportionate attention. To probe this further, I examined correlations with occupational prestige.

4.3.3 Prestige and Symbolic Status

Figure 4.4 shows fictional overrepresentation against survey-based prestige scores from Hughes *et al.* (2024). A positive trend is evident: higher-prestige occupations are more likely to be overrepresented (Spearman's $\rho = 0.40$, $p < 0.001$; results are substantively identical using Pearson correlation).

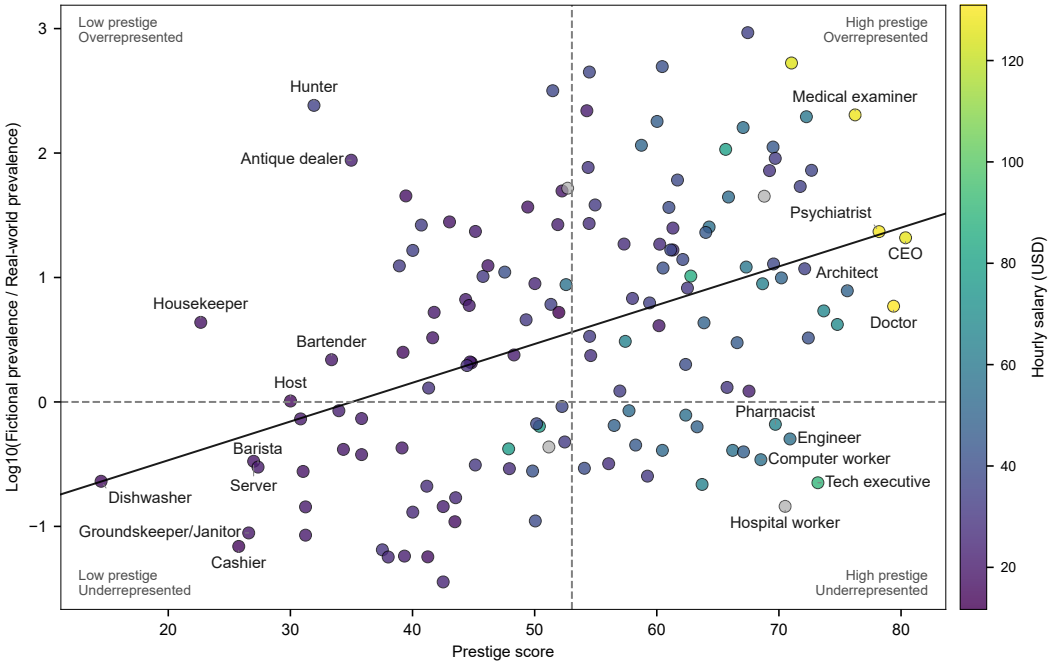

Figure 4.4.. Fictional representation of occupations vs. real-world occupational prestige. The y-axis shows the $\log_{10}$ ratio of fictional to real-world prevalence; the x-axis shows prestige scores from a survey (Hughes *et al.*, 2024). Points above the horizontal dashed line are overrepresented in fiction relative to their real-world frequency, while those below are underrepresented. The vertical dashed line marks the average prestige score across all occupations. Each point is colored by average real-world hourly wage (gray if unavailable). Together, the lines divide the space into four quadrants: high-prestige overrepresented, high-prestige underrepresented, low-prestige overrepresented, and low-prestige underrepresented.

Yet the distribution reveals key asymmetries. A large cluster of service and technical roles (e.g., cashier, server, barista, housekeeper) fall in the low-prestige underrepresented quadrant, underscoring fiction's neglect of the forms of labor that structure everyday life. Conversely, some low-prestige but narratively useful roles (e.g., hunter, antique dealer) are strongly overrepresented. Not all elite professions are equally elevated: doctors and CEOs are overrepresented compared to their prestige, while high-prestige, high-wage roles such as engineers and tech executives are underrepresented. These exceptions highlight how narrative utility can outweigh either economic reward or social prestige in determining which occupations gain prominence. In this sense, occupational prestige functions less as a proxy for earnings or education than as a cultural signal, shaping which roles fiction elevates or neglects.

While these broad patterns are informative, the genre of romance (and its offshoot fantasy romance) warrants closer attention, given its direct ties to aspirational and gendered portrayals of professions. Because these dynamics are especially pronounced and occupy a large share of the dataset (27% of books), I examine romance and fantasy romance in greater detail below.

4.3.4 Romance, Gender, and Desirability

To assess how occupational portrayals intersect with ideals of romantic desirability, I compare the ranked occupations of male and female protagonists in romance novels to real-life stated preferences from three sources (Table 4.2).[2] Two real-world data sources come from Tinder match rates for profiles listing different professions (Binder, 2016; Education Connection, 2022), and the third from a Zippia survey of career desirability (Morris, 2021). While differing in coverage and sampling, together they provide a useful point of comparison for how fictional desirability aligns—or fails to align—with real-world preferences.

Table 4.2 reveals sharp divergences. Real-world data consistently rank doctors, lawyers, and firefighters at the top for men—roles tied to stability, authority, and social respectability (Larson, 1977). Fiction, by contrast, elevates athletes, CEOs, and even crime bosses, foregrounding professions that symbolize power, risk, or transgression rather than occupational security (Messner, 1995; Ho,

[2] I excluded six same-sex romances due to insufficient sample size for separate analysis.

2009; Rafter, 2006). The prominence of male athletes is especially striking, with two-thirds of these being hockey players in particular. In contrast, professional athletes are rare in the U.S. (only 14 370 employed as of May 2024; U.S. Bureau of Labor Statistics, 2024). Their frequency in fiction illustrates how occupational desirability can function less as a reflection of labor markets than as a conduit of fantasy, attaching romance to professions culturally coded as embodying risk, masculinity, and fame.

Table 4.2.. Rankings of the most desirable professions for men (M) and women (W) in heterosexual relationships according to fictional portrayals and three real-world sources.

Rank (M/W)	Fictional ranking	Tinder sample 1	Tinder sample 2	Zippia survey
1	Athlete	Pilot	Firefighter	Doctor
2	CEO	Entrepreneur	Lawyer	Lawyer
3	Crime boss	Firefighter	Doctor	Carpenter
4	Musician	Doctor	Police officer	Engineer
5	Student	TV/radio host	Entrepreneur	Manager
6	Writer	Teacher	Pilot	Firefighter
1	Student	Physical therapist	Nurse	Nurse
2	Writer	Interior designer	Teacher	Elementary school teacher
3	Musician	Entrepreneur	Flight attendant	Doctor
4	Doctor	PR/communications	Server	Secretary
5	Artist	Teacher	Bartender	Lawyer
6	Florist	Student	Lawyer	Dancer

For women, the contrast takes a different shape. Fiction emphasizes roles tied to youth, creativity, and domestic aesthetics—student, writer, musician, artist, florist—while real-world rankings highlight care and service professions such as nurse, teacher, and secretary. The fictional absence of these caregiving roles, alongside the prominence of artistic and symbolic ones, underscores how desirability in women is often constructed through expressive rather than vocational identities (Ridgeway and Correll, 2004).

To further examine how these desirability scripts are distributed within romance genres, Figure 4.5 breaks occupations down by gender. In romance, the distribution echoes the pattern observed above: high-prevalence roles for men cluster around power, authority, or danger, while high-prevalence roles for women emphasize youth, domesticity, or support. Gender-neutral or cross-cutting roles are largely absent, underscoring how occupational identity is tightly bound to gendered desirability scripts.

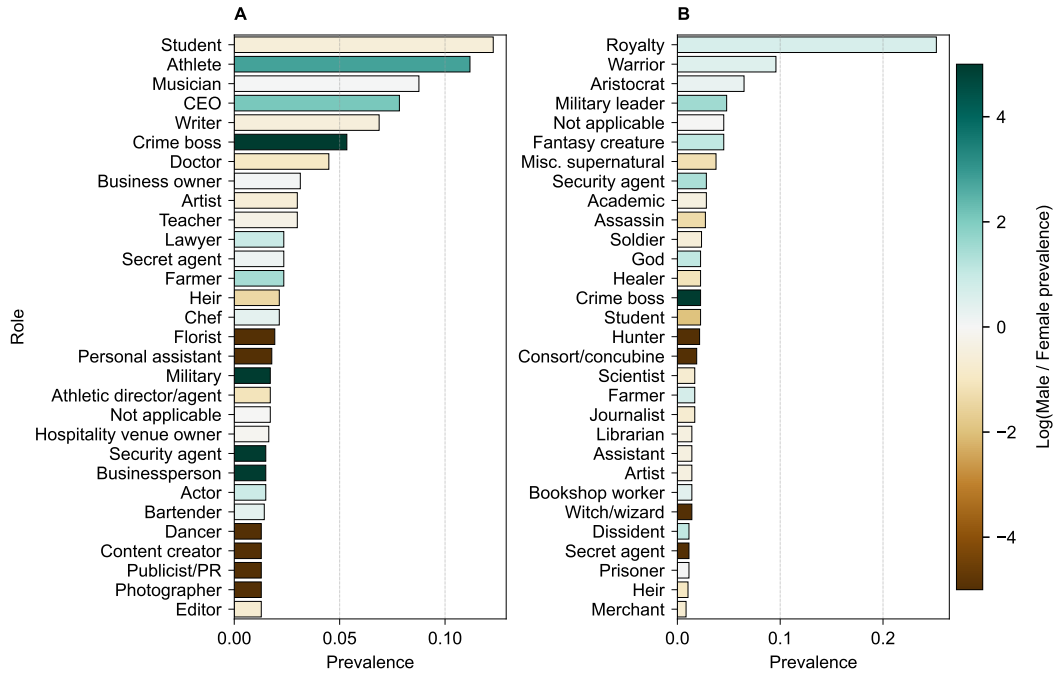

Figure 4.5.. Role prevalence in romantic fiction. Size of bars indicate prevalence of the role, and color indicates gender skew of characters, where higher and lower values indicate disproportionately male and female character's prevalence, respectively. Panel A: Romance. Panel B: Fantasy romance.

A different pattern emerges in fantasy romance. Roles such as warrior, royalty, healer, and witch/wizard introduce fantasy-world institutions that allow for more mixed gender associations. Some skews persist (e.g., warriors are predominantly men, healers more often women), but the overall distribution is less rigid. In this sense, romance codifies occupational identity into conventional gendered tropes, while fantasy romance destabilizes them, opening more space for variation beyond conventional gender sorting.

4.4 Discussion

This study demonstrates how fiction constructs symbolic hierarchies of work that diverge sharply from labor market realities. Across genres, a handful of archetypal roles—students, writers, detectives, athletes—dominate, while the vast majority of everyday occupations are absent. Prestige shapes representation to some degree, but it is not determinative: some elite roles (e.g., tech executives, engineers) fall below parity, while low-prestige but narratively useful roles (e.g., hunters, crime bosses) are foregrounded. Romance further

encodes occupational identity into gendered desirability scripts, with men disproportionately cast in roles of power or danger and women in roles of youth, creativity, or domesticity. Together, these findings highlight how occupational prestige is important less as a proxy for income than as a form of symbolic capital, shaping which roles are narratively elevated, sidelined, or erased.

The absence or underrepresentation of retail, service, and care work—occupations that employ large shares of the real population—is particularly telling. Their invisibility in popular fiction mirrors broader cultural dynamics of devaluation and erasure, long noted in studies of "invisible labor" and the feminization of care work (Glenn, 1992; England, 2005). Fictional portrayals thus reinforce symbolic boundaries between ordinary and exceptional labor: the former is pushed out of cultural visibility, while the latter is elevated as a marker of glamour, excitement, or desirability.

The gendered asymmetries in romance underscore how occupations operate not only as markers of class or status but also as vehicles for desire. Men are written into professions symbolizing risk, transgression, or authority, while women are attached to expressive or domestic identities rather than caregiving professions that dominate real-world rankings of women's desirability by profession. Because women are overrepresented among romance readers (Thelwall, 2019), these portrayals likely encode dual aspirational logics: the female character's role as relatable or aspirational for the reader, and the male character's role as desirable in a partner (Radway, 1991; Regis, 2003).

Methodologically, this study also contributes to the growing intersection of computational methods and cultural sociology. By combining large language model–assisted coding with manual validation, it demonstrates a scalable way to extract structured cultural information from narrative sources. The approach enabled the coding of more than a thousand novels at a resolution (profession-level data for protagonists and love interests) that would be prohibitive through manual reading alone. At the same time, the procedure carries limitations: models sometimes default to "unknown" rather than resolve ambiguous cases, and validation still required reading a significant number of texts. These limitations underscore the need for careful triangulation, but also point to the promise of hybrid approaches that combine machine scale with human oversight.

Several limitations warrant caution. In my analyses, I compare fiction with real-world U.S. baselines from recent years. While I excluded historical fiction, fantasy, and other non-contemporary genres, it remains possible that some novels in "realistic" genres are set in different times or places, particularly in works in translation.[3] Additionally, without readership demographics linked to specific titles, it is difficult to assess how these portrayals resonate across social groups. Some novels may reach broad, heterogeneous audiences, while others succeed within narrower niches.

Future research can extend this work in several directions. Longitudinal studies could examine how fictional occupations shift over time, tracing whether changes in labor markets (e.g., the rise of tech jobs) eventually register in cultural portrayals. Comparative work across media could test whether visual forms such as film or television lean more heavily on certain occupational archetypes. Cross-cultural analyses could contrast Anglo-American patterns with portrayals in non-English bestselling fiction. Finally, stronger real-world benchmarks would improve comparisons: while surveys and dating-app data provide useful proxies for occupational desirability, they remain limited. Dating-app samples are unrepresentative and often confounded by non-occupational traits, while existing prestige surveys capture only general esteem rather than context-specific judgments (e.g., for oneself, for a child, or for a partner).

In sum, this study demonstrates that popular fiction is not a neutral mirror of labor markets but a site where occupations are reframed as cultural symbols rather than economic roles. The absence of ordinary labor, the selective elevation of rare or narratively useful roles, and the gendered scripts of desirability together underscore how cultural products mirror and selectively amplify cultural hierarchies of work and worth. By combining computational scale with sociological interpretation, this analysis provides new evidence for how occupational identities are constructed and circulated in mass-market storytelling.

[3] I experimented with using GPT-4o to extract information about story settings in the same way I identified professions, but the model's performance was too poor to incorporate.

References

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). „Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk". In: *Political Analysis* 20.3, pp. 351–368.

Binder, Matt (2016). *These Are the Most Popular Jobs on Tinder*. `https://www.wired.com/story/tinder-popular-professions/`. Accessed: 2025-04-22.

Buss, David M. (1989). „Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures". In: *Behavioral and Brain Sciences* 12.1, pp. 1–14.

Cawelti, John G. (1976). *Adventure, Mystery, and Romance: Formula Stories as Art and Popular Culture*. Chicago: University of Chicago Press.

Coltrane, Scott and Michele Adams (1997). „Work-Family Imagery and Gender Stereotypes: Television and the Reproduction of Difference". In: *Journal of Vocational Behavior* 50.2, pp. 323–347.

Couldry, Nick (2012). *Media, Society, World: Social Theory and Digital Media Practice*. 1st. Cambridge: Polity.

Education Connection (2022). *Sexiest Careers for Men and Women*. `https://www.educationconnection.com/resources/sexiest-careers-for-men-and-women/`. Accessed: 2025-04-22.

England, Paula (2005). „Emerging Theories of Care Work". In: *Annual Review of Sociology* 31, pp. 381–399.

Glassdoor (2025). `https://www.glassdoor.com`. Accessed: 2025-04-22.

Glenn, Evelyn Nakano (1992). „From Servitude to Service Work: Historical Continuities in the Racial Division of Paid Reproductive Labor". In: *Signs: Journal of Women in Culture and Society* 18.1, pp. 1–43.

Ho, Karen (2009). *Liquidated: An Ethnography of Wall Street*. Duke University Press.

Holtzman, Linda and Leon Sharpe (2014). *Media Messages: What Film, Television, and Popular Music Teach Us About Race, Class, Gender, and Sexual Orientation*. 2nd. New York: Routledge, p. 558.

Hughes, Bradley T., Sanjay Srivastava, Magdalena Leszko, and David M. Condon (Feb. 2024). „Occupational Prestige: The Status Component of Socioeconomic Status". In: *Collabra: Psychology* 10.1, p. 92882.

Larson, Magali Sarfatti (1977). *The Rise of Professionalism: A Sociological Analysis*. 1st ed. University of California Press.

Messner, Michael A. (1995). *Power at Play: Sports and the Problem of Masculinity*. Boston: Beacon Press, p. 256.

Morris, Kathy (2021). *The Most (And Least) Attractive Jobs for a Romantic Partner*. https://www.zippia.com/advice/most-attractive-jobs-least/. Accessed: 2025-04-22.

Nelson, Laura K. (2020). „Computational Grounded Theory: A Methodological Framework“. In: *Sociological Methods & Research* 49.1, pp. 3–42.

Neuendorf, Kimberly A. (2017). „Reliability“. In: *The Content Analysis Guidebook*. 2nd ed. Chapter 6. SAGE Publications, Inc.

OpenAI (2024). *GPT-4o System Card*. Tech. rep. OpenAI.

Radway, Janice A. (1991). *Reading the Romance: Women, Patriarchy, and Popular Literature*. University of North Carolina Press.

Rafter, Nicole (Apr. 2006). „The Heroes of Crime Films“. In: *Shots in the Mirror: Crime Films and Society*. Oxford University Press.

Regis, Pamela (2003). *A Natural History of the Romance Novel*. University of Pennsylvania Press.

Ridgeway, Cecilia and Shelley Correll (Aug. 2004). „Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations“. In: *Gender & Society* 18, pp. 510–531.

Shackelford, Todd, David Schmitt, and David Buss (July 2005). „Universal Dimensions of Human Mate Preferences“. In: *Personality and Individual Differences* 39, pp. 447–458.

Thelwall, Mike (2019). „Reader and author gender and genre in Goodreads“. In: *Journal of Librarianship and Information Science* 51.2, pp. 403–430.

Treiman, Donald J. (1977). *Occupational Prestige in Comparative Perspective*. English. Quantitative studies in social relations. New York: Academic Press.

U.S. Bureau of Labor Statistics (2024). *Occupational Employment and Wage Statistics by Industry*. https://data.bls.gov/oes/#/industry/000000. Accessed: 2025-04-22.

Valentino, Lauren (2019). „What is a 'Good' Job? Cultural Logics of Occupational Prestige“. Retrieved from https://hdl.handle.net/10161/19827. Dissertation. Duke University.

Zippia (2025). https://https://www.zippia.com/. Accessed: 2025-04-22.

Zoonen, Liesbet van (1996). „Feminist Perspectives on the Media“. In: *Mass Media and Society*. Ed. by James Curran and Michael Gurevitch. 2nd. Hodder Education, pp. 31–52.

5 Assortative Marriage by Occupational Distance

Abstract

Studies of assortative marriage by occupation typically use a binary ”same occupation” indicator, which misses meaningful near-matches. I introduce a continuous measure of occupational proximity based on word2vec embeddings trained on Danish occupation sequences and apply it to all opposite-sex marriages formed in Denmark, 2011–2022. Observed couples are benchmarked against a shuffled counterfactual within marriage-month cohorts, and similarity is tracked in event time around marriage. Three metrics are compared: exact same-occupation (binary), ISCO-08 textual similarity (BERT-based), and word2vec similarity. Across cohorts, spouses are substantially more similar than chance, though the extent varies by metric. Similarity is elevated before marriage and remains above baseline afterward. Using restricted cubic splines within couple-matched conditional logit models, I find that marriage odds rise with occupational embedding similarity, and that this rise is steeper as similarity scores near 1. An increase from the 25th to the 75th percentile corresponds to about a 1.4× increase in odds, an effect which persists after adjusting for exact same-occupation matches. The results demonstrate that occupational homogamy is graded and provide a simple, portable measurement framework.

5.1 Introduction

Assortative marriage—the tendency for individuals to partner with others who share similar social, economic, or cultural traits—is a key explanatory mechanism of social organization (Luo, 2017). One long-standing indicator is *occupational* homogamy: spouses are more likely than chance to work in

the same occupation (Mansour and McKinnish, 2018). Most empirical work operationalizes this as a binary ”same occupation” match (yes/no), sometimes within a small set of canonical professions (e.g., doctors, lawyers) (Schwartz *et al.*, 2021). But in contemporary labor markets with thousands of distinct occupations, exact matches are rare and noisy; treating all non-matches as equally dissimilar discards substantial information about proximity between jobs that are functionally or socially adjacent (e.g., nurse vs. physiotherapist; software engineer vs. data analyst).

This paper argues that occupational homogamy is inherently continuous. Occupations vary in semantic and task-space proximity, and partners may be highly similar even without identical job titles. Advances in text- and network-based representations now permit constructing continuous measures of occupational similarity from task/skill content and co-occurrence patterns (see e.g., Djumalieva and Sleeman, 2018 or Chapter 3). Leveraging these tools, we can quantify the *degree* of similarity between partners' jobs and directly compare it to conventional same-occupation indicators.

The contribution is twofold. First, I develop and apply continuous measures of occupational proximity, derived from (i) the ISCO-08 hierarchy and (ii) a word-embedding representation of occupations, to reassess the magnitude of assortative marriage by occupation. Second, I leverage registry-linked timelines to trace how occupational proximity differs from a shuffled counterfactual in the years before and after marriage. While prior work has studied the evolution of assortativity around union formation (Oppenheimer, 1988), few studies have measured occupational similarity at a fine-grained level, and, to my knowledge, none have done so using embedding-based continuous metrics.

Denmark offers an ideal setting for this exercise. Comprehensive administrative registries link individuals' occupations to marriage and partnership histories at monthly resolution, enabling precise alignment of occupational trajectories with union formation. Using the universe of heterosexual marriages and registered partnerships formed between 2011 and 2022, I compare the observed similarity of spouses' occupations to a counterfactual obtained by randomly reassigning partners within marriage cohorts (holding cohort structure fixed). Similarity is assessed under three metrics: (1) exact occupational match (binary), (2) text description-based proximity within ISCO-08 (International

Labour Office, 2023), and (3) distance in a word2vec-derived embedding of occupations as described in Section 3.2.2.

Substantively, the approach speaks to two questions. *Measurement:* Do continuous similarity metrics reveal stronger or more stable occupational assortativity than binary same-job indicators, which may undercount meaningful near-matches? *Dynamics:* Is elevated similarity present primarily prior to marriage (selection into unions), or does it also change after marriage (adaptation via job moves or specialization)? Addressing these questions refines our understanding of how occupational structure contributes to household sorting and, by extension, to the distribution of resources and risks (Schwartz, 2013).

The results preview is as follows: (i) real couples are significantly more similar than shuffled pairs by each of the three metrics; (ii) similarity rises in the years leading up to marriage and remains elevated afterward, consistent with selection into unions documented in prior work (Kalmijn and Flap, 2001); and (iii) in logistic models of marriage formation, a continuous similarity measure predicts marriage even when controlling for a same-occupation indicator, providing evidence that the graded measure captures information beyond exact matches.

More broadly, the paper offers a measurement strategy for scholars of homogamy and social stratification—one that enables re-analyses of assortativity less dependent on rare exact matches and more sensitive to the structure of modern work. The remainder proceeds as follows: Section 5.2 introduces data and similarity measures; Section 5.3 presents benchmarking against shuffled pairs, marriage-probability plots by occupational similarity, and regression estimates; Section 5.4 discusses implications, limitations, and avenues for further work.

5.2 Materials and Methods

5.2.1 Data and Sample

I use Danish administrative registry data linking individuals' employment histories to vital events, including marriage, during 2011–2022 (Stender *et al.*,

2015). Couples are identified in the registries by the recorded date of marriage or registered partnership. The study population includes all opposite-sex couples who formed a legal union between 2011 and 2022 and for whom both partners have at least one observed employment spell in 2011–2022.[1]

Employment states are measured using national employment registers, with occupations recorded at a monthly level of granularity using Danish six-digit occupation codes (Statistics Denmark, 2011), the first four digits of which are the ISCO-08 code (International Labour Office, 2023), and the last two digits of which are a potential refinement. For each individual and month t, I define the person's primary occupation as the one at which they worked the most hours. Event time is indexed in months relative to the union date, $\tau \in \{-143, \ldots, +143\}$, with $\tau = 0$ defined as the calendar month of marriage.

5.2.2 Occupational Similarity Measures

I assess partner similarity using three metrics that move from categorical to continuous proximity:

(1) Same-occupation (binary). A couple is coded as similar at time τ if partners share the same six-digit occupation code (1 if identical, 0 otherwise). This replicates the conventional operationalization of occupational homogamy.

(2) ISCO-based textual similarity. To capture semantic proximity within the ISCO-08 classification, I compute similarity directly from official occupational descriptions. Each ISCO-08 code has a standardized textual description specifying core tasks and responsibilities. I embed these texts using a pre-trained BERT model (Reimers and Gurevych, 2019), yielding contextualized sentence vectors for each occupation. The similarity between partners' occupations $c_{i\tau}$ and $c_{j\tau}$ at event time τ is then defined as the cosine similarity between their respective embeddings:

$$\text{ISCO-Sim}_{ij\tau} \;=\; \cos\Big(\mathbf{b}_{c_{i\tau}},\, \mathbf{b}_{c_{j\tau}}\Big) = \frac{\mathbf{b}_{c_{i\tau}}^{\top}\mathbf{b}_{c_{j\tau}}}{\|\mathbf{b}_{c_{i\tau}}\|_2\,\|\mathbf{b}_{c_{j\tau}}\|_2} \;\in [-1, 1],$$

[1] Same-sex couples are excluded for conceptual comparability with prior work and because the distribution of occupations and matching processes may differ in ways that merit a separate analysis (Schwartz and Graf, 2009).

where $\mathbf{b}_c$ denotes the BERT-derived embedding of the ISCO-08 description for occupation c. Higher values indicate greater semantic similarity between the textual content of the occupational definitions, and thus a closer alignment in underlying task and skill structure.

(3) Embedding (word2vec) similarity. I trained a word2vec skip-gram model on Danish job trajectory data to obtain d-dimensional occupation vectors $\mathbf{v}_c \in \mathbb{R}^d$ (here $d = 100$). Vectors are L^2-normalized, and similarity is the cosine:

$$\text{EmbedSim}_{ij\tau} = \cos\left(\mathbf{v}_{c_{i\tau}}, \mathbf{v}_{c_{j\tau}}\right) = \frac{\mathbf{v}_{c_{i\tau}}^{\top} \mathbf{v}_{c_{j\tau}}}{\|\mathbf{v}_{c_{i\tau}}\|_2 \, \|\mathbf{v}_{c_{j\tau}}\|_2} \in [-1, 1].$$

Higher values indicate greater proximity in the learned job embedding space. The word2vec model was trained with a window size of 7, a minimum token prevalence of 1, a negative sampling rate of 10^{-3}, for 95 epochs. For more details about the word2vec model, see Chapter 3 and Section 2.2.4.

Occupation codes are missing for individuals who aren't employed wage-earners. In these cases, missing occupation codes yield SameOcc $= 0$.

5.2.3 Counterfactual Construction

To benchmark observed similarity, I construct a counterfactual by randomly reassigning partners within marriage cohorts. For each observed couple (i, j) married in month and year (m, y), I draw K shuffled partners j' from the pool of individuals of the opposite sex who marry in (m, y), and form pseudo-couples (i, j'). Similarity is computed for each pseudo-couple at each event time τ. This design preserves cohort structure and the event-time alignment while breaking the within-couple occupational link.

5.2.4 Event-time Estimands

Let $\text{sim}_{ij\tau}$ denote a given occupation-similarity metric for couple (i, j) at event time τ, where $\tau = 0$ marks the month of marriage. For each metric, I estimate

the average difference in similarity between observed couples and a cohort-preserving shuffled baseline:

$$\Delta(\tau) = \mathbb{E}[\text{sim}_{ij\tau} \mid \text{real}] - \mathbb{E}[\text{sim}_{ij\tau} \mid \text{shuffled}]\,.$$

$\Delta(\tau)$ represents the absolute lift in similarity relative to the shuffled benchmark, with higher values indicating stronger assortative sorting. To test whether the observed lift differs from zero, I obtain a 95% confidence band from the empirical distribution of shuffled-similarity deviations (a replicate minus the across-replicate mean) over 1,000 shuffled replicates, which serves as the null distribution under $\mathbb{E}[\Delta(\tau)] = 0$.

5.2.5 Marriage Propensity as a Function of Similarity

To visualize how the probability of marriage varies with occupational similarity, I group all potential dyads into bins of embedding similarity (width 0.067). For each bin b, I compute the relative probability of marriage, which I denote by *enrichment*, as

$$E(b) \;=\; \frac{P(\text{married} \mid \text{similarity bin } b)}{P(\text{married overall})},$$

where the numerator is the share of real couples whose similarity falls in bin b, and the denominator is the overall probability of marriage among all possible dyads in the same monthly cohort. Values of $E(b) > 1$ therefore indicate enrichment-bins where marriage occurs more often than expected given baseline probability of marriage. I plot $E(b)$ against the mean similarity in each bin. This nonparametric enrichment measure provides a straightforward visualization of how occupational proximity relates to marriage formation without imposing any functional form or model-based assumptions.

5.2.6 Conditional Logit Models

To quantify how occupational similarity relates to marriage formation, I estimate *couple-matched conditional logit* models. For each observed marriage between partners (i, j) in month t, I construct a matched choice set (stratum) consisting of the real couple ($Y = 1$) and $K = 10$ randomly paired pseudo-

couples ($Y = 0$) formed by re-matching all men and women who marry in the same calendar month t. The conditional logit therefore compares each real couple only to alternative partners available in the same marriage month, holding fixed all month-specific conditions.

Because the conditional likelihood conditions out the stratum-specific intercept, the models take the form:

$$\begin{aligned}
\text{(M1)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_1\, \text{SameOcc}_{ij}, \\
\text{(M2)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_2\, \text{EmbedSim}^*_{ij} + \sum_{r=1}^{2} \gamma_r\, \text{RCS}_r(\text{EmbedSim}^*_{ij}), \\
\text{(M3)}\ \operatorname{logit}\Pr(Y_{ij} = 1 \mid s) &= \beta_1\, \text{SameOcc}_{ij} + \beta_2\, \text{EmbedSim}^*_{ij} \\
&\quad + \sum_{r=1}^{2} \gamma_r\, \text{RCS}_r(\text{EmbedSim}^*_{ij}),
\end{aligned}$$

where EmbedSim^*_{ij} is the standardized version of EmbedSim_{sj}, and $\text{RCS}(\text{EmbedSim}^*_{ij})$ is a restricted cubic spline basis (with $k = 4$ knots, yielding $k - 2 = 2$ spline terms). The spline knots are placed at empirical quantiles of EmbedSim^*, and all spline components are evaluated on the standardized scale.

I report odds ratios for the exact-occupation indicator, and interpretable contrasts in embedding similarity (e.g., the increase in odds when moving from the 25$^{\text{th}}$ to the 75$^{\text{th}}$ percentile of similarity).

Standard errors are clustered by marriage month, the level at which matched sets are constructed. No additional covariates are included: demographic and socioeconomic attributes (age, education, occupation structure, labor market position) are part of the mechanism that the similarity measures are designed to summarize.

5.3 Results

5.3.1 Trajectories of Assortative Similarity

Figure 5.1 plots the difference of observed and randomized occupational similarity as a function of years before and after marriage. Values significantly

above zero indicate that couples are more similar in occupation than expected under random pairing.

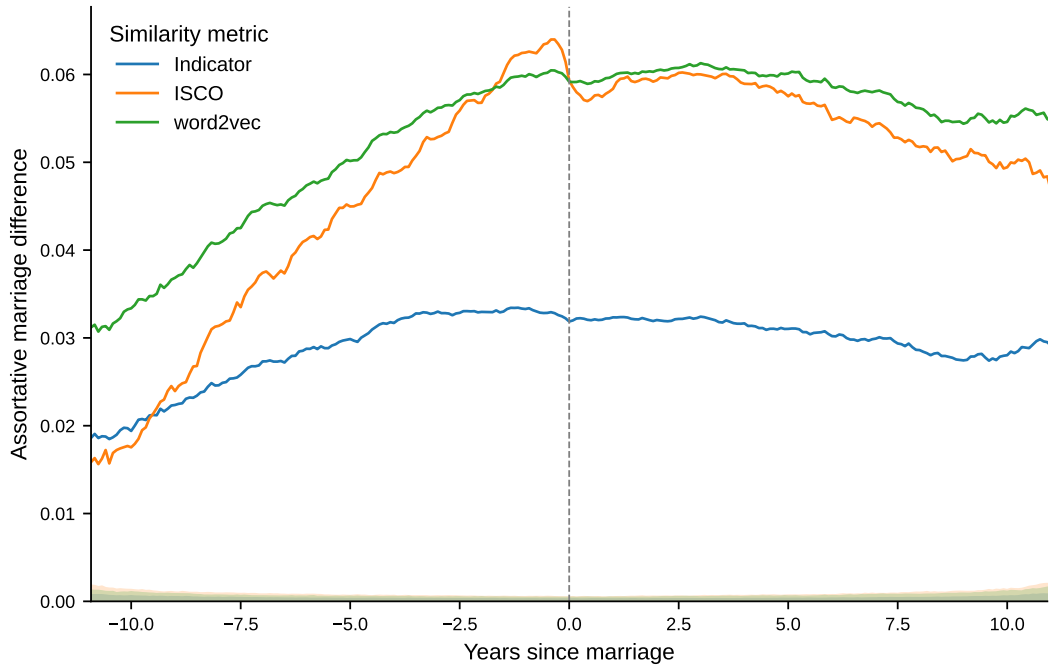

Figure 5.1.. Assortative occupational similarity around marriage. The figure shows the difference of observed to randomized occupational similarity by years relative to marriage ($\tau = 0$, indicated by a vertical dashed gray line). Values > 0 occur when spouses are more similar than expected under random pairing within the same marriage cohort. Three measures are plotted: a binary same-occupation indicator, ISCO similarity (BERT-based cosine similarity of ISCO-08 descriptions), and word2vec embedding similarity. 95% confidence bands around zero indicate the similarity expected from sampling couples randomly.

Across all years before and after marriage and by all three metrics, spouses are consistently more similar than chance would predict. The word2vec and ISCO differences are highest overall, with the ISCO indicator having a steeper climb before marriage, and a faster drop-off afterwards. This provides evidence that continuous similarity captures spousal similarity beyond exact matches.

The timing pattern is similar in broad outline—elevations are already present before marriage and persist afterward—suggesting that much of the alignment reflects selection into unions rather than post-marital convergence, with only modest scope for co-movement thereafter.

In subsequent cross-sectional analyses, I rely on metrics measured at the month of marriage. I select just one of the continuous metrics, the word2vec metric,

for demonstration purposes. Similar analyses could be conducted using ISCO similarity instead; see Figure C.1 for a plot comparing the two metrics at month of marriage.

5.3.2 Marriage Enrichment by Continuous Similarity

To illustrate the informational gain from a continuous metric, I plot an enrichment curve using word2vec similarity at the month of marriage. Unlike the binary same-occupation indicator, which yields a single point at similarity $= 1$, the embedding permits a full profile over near matches.

For each marriage cohort (year–month), I compare the number of observed couples in each similarity bin to the number of possible male-female pairings in that cohort, then aggregate across cohorts. For this cross-sectional view at marriage, couples with missing wage-earner occupation codes are excluded to avoid conflating dissimilar jobs with no recorded wage-earner job at marriage for at least one partner (43% of couples fall into the latter category).

Figure 5.2 shows a strongly right-skewed pattern: the probability of marriage rises with occupational similarity. Enrichment is near baseline (≈ 1) for pairs with similarity around 0.2, dips below one for negative similarity (dissimilar occupations), and then rises steadily with similarity. By mid–range similarity (roughly 0.4–0.6), enrichment is about two to three times the baseline. At the highest similarity levels ($\gtrsim 0.9$), enrichment peaks around eight to nine times the baseline.

5.3.3 Logistic Regressions of Marriage Likelihood

To quantify how occupational similarity predicts marriage formation, I estimate a series of conditional logistic regressions in which each observed couple constitutes its own choice set. For every real marriage $(Y = 1)$, I construct $K = 10$ counterfactual dyads $(Y = 0)$ by randomly re-pairing couples who marry in the same calendar month.

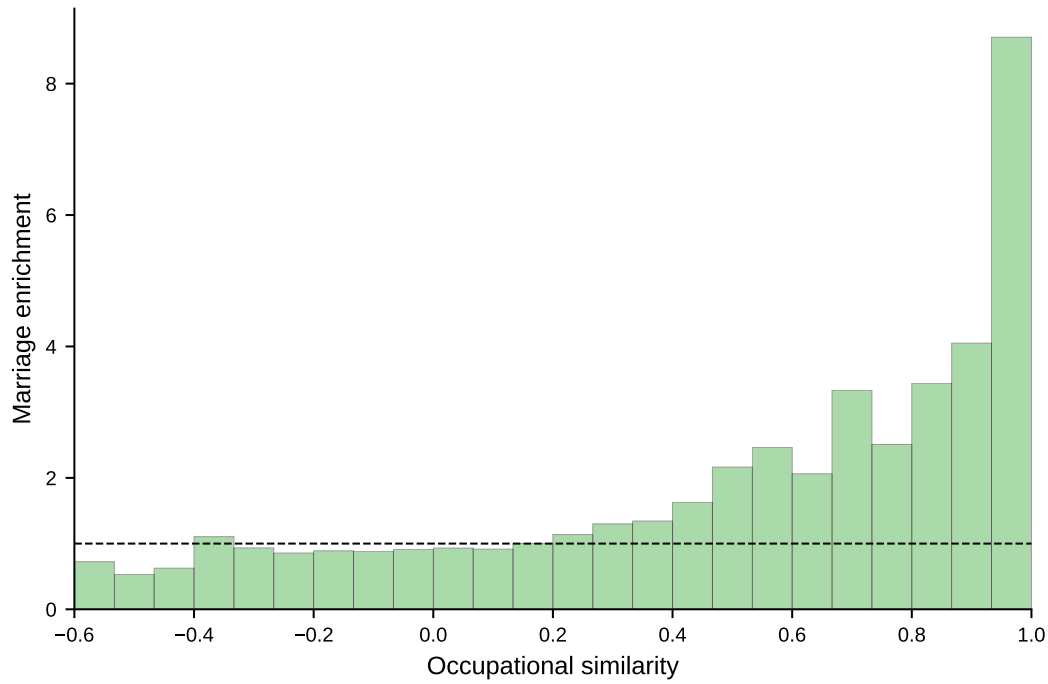

Figure 5.2.. Relative likelihood of marriage by word2vec occupational similarity. A bar chart of relative likelihood of marriage as a function of word2vec occupational similarity at the month of marriage. The dashed red horizontal line at 1 indicates baseline likelihood. Couples where at least one partner has a missing wage-earner occupation code at marriage are excluded from this figure.

Table 5.1 reports odds ratios for the three models described in Section 5.2.6.

Table 5.1.. Conditional logit models of marriage on occupational similarity (odds ratios).

	(M1) SameOcc only	(M2) RCS(EmbedSim)	(M3) SameOcc + RCS(EmbedSim)
Same occupation			
OR (95% CI)	9.23 (9.00, 9.46)	–	0.78 (0.73, 0.83)
Embedding similarity			
OR: P25 → P75	–	1.41	1.40
OR: P10 → P90	–	2.94	3.04
OR: P80 → P95	–	7.16	8.86

Notes: Odds ratios (OR) from conditional logit models matching each observed couple to $K = 10$ randomly sampled counterfactual pairings within the same marriage month. Models (M2) and (M3) use a restricted cubic spline (RCS) for the continuous embedding similarity measure. Reported contrasts correspond to changes in similarity between empirical percentiles of the real-couple distribution. Standard errors are cluster-robust by marriage month. Full coefficient tables are provided in Appendix C.2.
OR confidence intervals are based on asymptotic standard errors from the conditional logit. Because pseudo-couples are constructed by sampling men and women from the same marriage-month pool, observations are not fully independent: selecting an individual into one pseudo-pair affects the availability of that individual for other pairs. As a result, standard errors may be optimistic, and the reported CIs should be interpreted as approximate.

The exact-match indicator is strongly associated with marriage: sharing an identical occupation is linked to an odds ratio of approximately 9.23 (95% CI: 9.00, 9.46). This replicates prior findings that same-occupation couples are substantially overrepresented relative to chance.

When modeling embedding similarity flexibly using a restricted cubic spline with four knots, the association between similarity and marriage odds remains strong and highly significant. Because the spline is nonlinear and coefficients are not directly interpretable, I summarize model-implied contrasts on the original similarity scale. Moving from the 25th to the 75th percentile of embedding similarity increases the odds of marriage by approximately 1.41×; moving from the 10th to the 90th percentile yields an increase of 2.94×. At the upper tail, a contrast from the 80th to 95th percentile corresponds to 7.16×.

Including both predictors simultaneously allows us to assess whether the continuous similarity measure captures information beyond exact matches. In this analysis, the spline-based similarity effect remains strong: the $25 \rightarrow 75$ percentile contrast is 1.40, and the $10 \rightarrow 90$ percentile contrast is 3.04. Meanwhile, the same-occupation odds ratio attenuates sharply to 0.78 (95% CI: 0.73, 0.83). The attenuation is expected: continuous embedding similarity assigns

the value 1 to exact matches, so a substantial portion of the same-occupation effect is absorbed by the nonlinear spline.

The restricted cubic spline specification was selected based on a formal likelihood-ratio test, which strongly rejected a purely linear effect of embedding similarity ($p < 0.001$). The conditional c-index and top-1 accuracy are 0.59 and 0.20 for both (M2) and (M3), while they are 0.53 and 0.14 for (M1), further indicating that the model is strengthened by incorporating a continuous similarity term.

Overall, the results reinforce the enrichment and event-time analyses: occupational sorting in marriage reflects both broad affinities encoded continuously and strong preferences for exact occupational matches.

5.4 Discussion

This paper reframes occupational homogamy as a continuous phenomenon. Using Danish registry data linked to detailed occupations, I show that spouses are far more similar in occupation than random pairing would imply, and that a word2vec-based metric yields a smooth gradient of proximity which is associated with marriage formation even when exact same-occupation matches are taken into account. The event-time profiles indicate that elevated similarity is already present prior to marriage and remains above the shuffled baseline afterward, consistent with assortative selection as the primary driver, with limited scope for post-marital convergence. Furthermore, the word2vec approach performs better than an ISCO textual similarity benchmark, suggesting that embeddings trained on labor-market co-occurrences capture task/skill proximity not fully reflected in taxonomy descriptions.

Several design choices constrain interpretation. The longitudinal trajectories are aligned to the marriage month; as time elapses after marriage, some unions dissolve or are censored by death, and I do not adjust for differential exposure at later τ. The shuffled counterfactual design also implicitly assumes that occupational trajectories would have been unchanged under different partner assignments; if spouses influence one another's careers through relocation, childcare, or informal job search assistance, this assumption may understate

genuine post-marital adaptation. In the longitudinal analyses, months without wage-earner occupation codes were assigned similarity 0 and, for the cross-sectional enrichment plot at marriage, such cases were excluded. This convention avoids conditioning on continuous dual employment but attenuates average levels; the real-versus-shuffled comparison remains valid because the rule is applied symmetrically. Shuffling within marriage month approximates but does not perfectly capture the opportunity set of potential partners. Finally, the Danish institutional context may limit generalizability.

Several straightforward extensions could deepen or broaden the approach. Earlier relationship markers—first cohabitation or first birth of a child—could be used where available to probe similarity at union initiation for couples who marry later or never marry. This is especially relevant in the Danish context where couples may cohabitate for years before marriage, or forego legal marriage altogether (Kasearu and Kutsar, 2011). Additional similarity spaces could be compared against word2vec, including skills-based distances such as O*NET task/skill vectors and alternative embedding architectures (Djumalieva and Sleeman, 2018). Coverage could be expanded to self-employed and assisting spouses via industry codes and owner registers, and the state space could be extended to education, unemployment, and out-of-labor-force states. It would also be informative to track changes by calendar time and to examine heterogeneity by education, age differences, or urbanicity, and to extend the analysis to same-sex couples.

In conclusion, treating occupation as a point in a continuous similarity space reveals a graded structure of assortative marriage that binary same-occupation indicators necessarily miss. A simple embedding-based measure provides a sharp, scalable lens on occupational proximity and aligns with intuitive enrichment and time-series diagnostics. As a methodological note, the takeaway is straightforward: replacing ”same job” with a continuous occupational similarity can substantially improve the descriptive accuracy of assortative-marriage research, with minimal additional complexity.

References

Djumalieva, Jyldyz and Cath Sleeman (Jan. 2018). „An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts“. In: pp. 425–454.

International Labour Office (2023). *The International Standard Classification of Occupations (ISCO-08) Companion Guide*. Geneva: International Labour Office.

Kalmijn, Matthijs and Henk Flap (2001). „Assortative Meeting and Mating: Unintended Consequences of Organized Settings for Partner Choices“. In: *Social Forces* 79, pp. 1289–1312.

Kasearu, Kairi and Dagmar Kutsar (May 2011). „Patterns Behind Unmarried Cohabitation Trends in Europe“. In: *European Societies* 13.2, pp. 307–325.

Luo, S. (2017). „Assortative mating and couple similarity: Patterns, mechanisms, and consequences“. In: *Social and Personality Psychology Compass* 11.8, e12337.

Mansour, Hani and Terra McKinnish (2018). „Same-occupation spouses: preferences or search costs?“ In: *Journal of Population Economics* 31, pp. 1005–1033.

Oppenheimer, Valerie Kincade (1988). „A Theory of Marriage Timing“. In: *American Journal of Sociology* 94.3, pp. 563–591.

Reimers, Nils and Iryna Gurevych (Nov. 2019). „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks“. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schwartz, Christine and Nikki Graf (2009). „Assortative matching among same-sex and different-sex couples in the United States, 1990-2000“. In: *Demographic Research* 21.28, pp. 843–878.

Schwartz, Christine, Yu Wang, and Robert Mare (Aug. 2021). „Opportunity and change in occupational assortative mating“. In: *Social Science Research* 99, p. 102600.

Schwartz, Christine R. (2013). „Trends and Variation in Assortative Mating: Causes and Consequences“. In: *Annual Review of Sociology* 39.Volume 39, 2013, pp. 451–470.

Statistics Denmark (Mar. 2011). *DISCO-08: Danmarks Statistiks fagklassifikation*. Technical manual. First edition. Statistics Denmark.

Stender, Pernille, Thomas Thorsen, and Hans Henrik Andersen (2015). „Micro data integration for Labour Market Account“. In: *Statistical Journal of the IAOS*.

6 Discussion

6.1 Introduction

The purpose of this chapter is to synthesize the findings from the three projects presented in this dissertation and to reflect on their broader theoretical, methodological, and empirical implications. While each project addressed a distinct research question—ranging from the structural organization of the labor market to the symbolic representation of work in fiction and the dynamics of assortative marriage—all share a common analytical foundation. Each treats occupations as key sites where structure and meaning intersect, and each employs computational methods to make these social dynamics empirically visible at scale.

The discussion proceeds in six sections. Section 6.2 examines how the first and third projects reconceptualize occupations as *structural units* in the labor market. Section 6.3 then turns to occupations as *symbolic units*, drawing primarily on findings from Chapter 4 to explore how work functions as a cultural and narrative device in contemporary fiction. Section 6.4 reflects on the role of computational methods in social sciences, assessing both their promise and their epistemological limits.

Section 6.5 considers the strengths and limitations of the dissertation as a whole, including the trade-offs inherent in working with administrative data and automated text analysis. Section 6.6 outlines directions for future inquiry, both substantive and methodological, arising from the findings. The chapter concludes in Section 6.7 by returning to the dissertation's central theme: that occupations occupy a dual role in modern societies, serving simultaneously as the building blocks of economic structure and as the symbolic vocabulary through which identities, aspirations, and hierarchies are expressed.

6.2 Occupations as Structural Units

Occupations have long been central to the analysis of social stratification, mobility, and inequality. In classical sociology, they serve as proxies for access to resources and as markers of social class. In this dissertation, the first and third projects revisit this tradition using computational methods to model occupations not as static categories but as dynamic, relational positions within the labor market. The embedding-based framework developed in Chapter 3 redefines occupational boundaries from the bottom up, grounding similarity in observed career transitions rather than in expert taxonomies. Chapter 5 extends this logic, using distance-based measures to reexamine assortative marriage.

This approach has several strengths. First, it uses job similarity from observed transitions rather than from abstract classification criteria, making it sensitive to actual labor market mobility patterns. Second, it flexibly incorporates employment states that are often marginalized or inconsistently handled in traditional models—part-time work, self-employment, unemployment, and student jobs—thereby yielding a more inclusive view of the labor market. Third, the method scales efficiently to the population-level Danish registry data, providing monthly granularity across more than a decade. This enables the analysis of structural shocks, such as the COVID-19 pandemic, with temporal precision rarely available in cross-sectional or survey-based research.

Beyond methodological innovation, the findings contribute to debates on the permeability of occupational boundaries (Gunz *et al.*, 2007). The 311 data-driven clusters derived in Chapter 3 reveal both expected groupings—such as coherent professional domains linked by education—and unexpected connections that cut across formal classifications, reflecting real-world hybridization of work. Such clusters can be interpreted as ”meso-level” units of labor market organization: more granular than class schemas but more interpretable than the thousands of individual occupations used in administrative coding.

Chapter 5 further demonstrates how these embeddings can refine classical measures of assortative marriage in the context of occupations. By replacing a binary ”same-occupation” indicator with a continuous, data-driven similarity measure, it becomes possible to detect graded patterns of assortative marriage

that are invisible in categorical approaches. Couples tend to pair not only within the same occupation but also across occupations that share mobility pathways. This relational perspective extends theories that tie social proximity and network closure to the reproduction of stratification, grounding them in high-resolution empirical data.

Nevertheless, several limitations remain. The embedding-based approach is inherently descriptive: it maps transitions as they occur, without disentangling structural constraints from individual preferences. Moreover, while Denmark's registry infrastructure provides unparalleled data coverage, its institutional context—characterized by flexicurity, high female labor force participation, and a compressed wage distribution—differs from that of many other countries. Some patterns, such as the degree of occupational fluidity or gendered specialization, may therefore be context-specific. Yet the underlying logic of the approach is broadly generalizable.

In sum, treating occupations as dynamic structural units highlights both continuity and change in the organization of work. Computational models enable us to visualize and quantify the labor market's hidden architecture, revealing the pathways through which inequality is produced, maintained, and occasionally reconfigured.

6.3 Occupations as Symbolic Units

Occupations are not only positions within the economic structure but also symbolic resources through which identity, aspiration, and value are expressed. The second project extends the analysis of work into the cultural domain, examining how fiction encodes and reproduces hierarchies of labor. By analyzing over a thousand bestselling novels, it reveals how cultural products translate economic categories into symbolic hierarchies that both mirror and distort real labor market structures.

Fiction constructs a world of work that is sharply selective. Across genres, a narrow set of archetypal roles occupy a disproportionate share of narrative attention, while the vast majority of everyday occupations are absent. This skew suggests that occupational prestige functions not merely as a reflection

of social standing but as a form of symbolic capital that governs which kinds of work are narratively salient. Yet prestige is not the sole determinant of visibility. Some prestigious professions in real life are underrepresented, while others with limited real-world presence flourish as vehicles for character development. Fiction thus operates as a symbolic economy of labor, in which narrative usefulness, emotional resonance, and genre convention outweigh statistical or economic realism.

The underrepresentation of retail, service, and care work is particularly revealing. These roles employ a large share of the real population but remain largely invisible in popular fiction, reflecting broader patterns of cultural devaluation (England, 2005; Glenn, 1992). In this sense, fiction does not simply entertain—it participates in boundary-making, delineating which forms of labor are worthy of attention and which remain unspoken.

The gendered patterning of occupational roles in romance further underscores how work functions as a shorthand in conveying desire. Men are disproportionately written into professions symbolizing danger, authority, or control, whereas women are associated with creative, nurturing, or domestic roles. These symbolic asymmetries map onto enduring cultural scripts of masculinity and femininity. Because women constitute the majority of readers in romance and related genres, these portrayals encode both identification and aspiration: female protagonists may reflect readers' ambitions or emotional worlds, while male characters embody an idealized partner.

These findings have broader implications for how we interpret occupations as symbolic resources. In fiction, work is not merely background detail but a semiotic shorthand through which social identity is established, moral character is inferred, and romantic compatibility is created. The same logic underlies, in more subtle form, real-world processes of partner selection and social perception, linking this project conceptually to the assortative marriage study—both suggest how occupations operate as social signifiers.

That said, these conclusions are necessarily bounded by scope. The corpus analyzed here comprises U.S. bestsellers, which reflect the tastes of a culturally specific readership. Yet their reach means they help define what kinds of labor are seen, celebrated, or ignored in contemporary culture. As such, they

illuminate not only the structure of narrative imagination but also the means through which societies make sense of work.

6.4 Computational Methods in Sociology

The three projects together illustrate how computational methods can extend the empirical reach of sociological inquiry while remaining grounded in theoretical concerns. By leveraging techniques such as embeddings, clustering, and large language model (LLM)–assisted text analysis, the dissertation demonstrates how computational tools can reveal latent patterns of structure and meaning that are inaccessible to traditional small-N or survey-based approaches. At the same time, these methods introduce distinctive challenges of interpretation, reproducibility, and ethical accountability that must be explicitly confronted.

In the structural analyses (Chapters 3 and 5), embedding models and clustering techniques made it possible to represent the labor market as a high-dimensional relational space, where occupations are positioned according to observed mobility flows. This approach yields an empirically grounded, data-driven classification system that captures the fluidity of modern work more effectively than static taxonomies. Clustering methods translate this space into interpretable units, while random recombination baselines serve as statistical counterfactuals, clarifying what aspects of social structure arise from genuine organization rather than from chance. Together, these methods demonstrate the analytical power of computational modeling when coupled with social scientific reasoning about structure, mobility, and closure.

In the cultural analysis (Chapter 4), computational methods enabled a complementary extension into symbolic domains. By combining LLM-assisted extraction with human validation, the project operationalized cultural representation at scale, systematically identifying occupational archetypes across more than a thousand novels. This hybrid workflow reflects a productive methodological synthesis: automation provides breadth and consistency, while human review preserves interpretive validity. It represents one model for computational cultural sociology that is neither fully automated nor wholly manual, but instead grounded in both machine inference and human judgment.

Despite their power, these tools come with interpretive limits. Embeddings capture associations in data but do not by themselves explain the social processes that produce them. LLMs, meanwhile, encode linguistic and cultural priors that may result in systematic errors, requiring transparency in prompting and validation. Moreover, computational methods often trade interpretability for scalability. The challenge is thus to use these models as heuristic devices—not as replacements for theory, but as instruments for identifying patterns that call for further interpretation.

6.5 Strengths and Limitations of the Dissertation

This dissertation's strengths lie in its combination of large-scale empirical data and cross-disciplinary scope. By integrating population-level administrative registries with large cultural text corpora, it bridges two major domains of sociological inquiry—the structural and the symbolic. The use of computational methods, from embeddings and clustering to LLM-assisted coding, demonstrates how new tools can be used to revisit long-standing questions about work, mobility, and meaning. Together, the three projects show that it is possible to link fine-grained labor market data with cultural representations at scale, revealing both how occupations are organized in practice and how they are imagined in narrative form.

The dissertation benefits from the unique advantages of the Danish data environment. The combination of universal registry coverage, longitudinal granularity, and individual-level linkage enables analyses that would be infeasible in most other contexts. These features allow for a rare population-level analyses, free from sampling error and attrition.

Several limitations nonetheless warrant caution. The temporal window (2011–2022) used in Chapters 3 and 5 captures only a single decade, during which major labor market disruptions—most notably the COVID-19 pandemic—occurred. Because the embedding model aggregates transition behavior across this period, it may obscure temporal variation in mobility. In addition, measurement inconsistencies in occupational labeling and income, particularly

among the self-employed and co-working spouses, introduce some noise into the analysis.

For the cultural corpus, limitations are partly conceptual and partly technical. The analysis focuses on English-language bestsellers from 2023–2024, a commercially defined and culturally specific subset of global fiction. Although non-contemporary or fantastical settings were excluded where possible in real-world labor market comparisons, some novels labeled as ”realistic” may nonetheless take place in other times or locations. Automated genre and setting detection, while tested, proved insufficiently accurate for inclusion. Moreover, the lack of readership demographic data limits interpretation of audience reach and reception.

Finally, the project presented in Chapter 5 remains exploratory. It illustrates the potential of embedding-based distance measures for family and stratification research but should be seen as a methodological proof of concept rather than a definitive analysis. Future work could integrate demographic covariates, temporal dynamics, or cross-national comparisons to build on this foundation.

In sum, the dissertation’s strengths lie in its scale, integration, and methodological ambition, while its limitations stem from the very innovations that make it possible: dependence on novel computational tools, reliance on high-context national data, and the interpretive challenges of connecting structure to meaning across distinct empirical domains.

6.6 Future Research Directions

This dissertation opens several avenues for future research at the intersection of labor market sociology, cultural analysis, and computational methodology. The three projects presented here demonstrate how structural and symbolic analyses of occupations can be brought into the same empirical framework; extending this work will require both broader datasets and more dynamic modeling techniques.

Extending structural analyses. Future research can build on the coarse-graining framework developed in Chapter 3 by applying it to other countries and institutional settings. Comparative studies across welfare regimes or labor market systems would help assess the generalizability of data-driven occupational clustering, clarifying whether the Danish pattern of high mobility and a compressed wage distribution produces distinctive community structures. Another promising direction is the use of dynamic coarse-graining methods to capture how occupational structures evolve in response to technological change and policy reforms. Linking these dynamic models to worker outcomes, such as wage growth or job security, would deepen our understanding of how labor markets evolve. As for assortative marriage, extending matching from occupational similarity to broader similarity across the lifespan using embeddings trained on life events is a promising avenue for further work.

Extending cultural analyses. The second project highlights the potential of computational cultural sociology but also points to clear opportunities for expansion. Longitudinal corpora could track how occupational portrayals shift over decades, testing whether cultural representations respond to real-world labor trends such as the rise of digital and care economies. Cross-media comparisons—across novels, film, and television—could reveal how different storytelling formats rely on distinct occupational archetypes. Cross-linguistic and cross-cultural analyses would further uncover whether the representations of work identified here are uniquely Anglo-American or part of a broader global culture. Finally, improving real-world comparison datasets remains a priority: prestige surveys, dating-app samples, and occupational desirability rankings each capture only partial aspects of social valuation. Future research might combine multiple indicators or develop new instruments that measure occupational prestige across relational contexts (e.g., ”for oneself,” ”for one’s child,” ”in a romantic partner”).

6.7 Conclusion of the Discussion

This Discussion chapter has drawn together the structural, symbolic, and methodological strands of the dissertation to address its overarching research question: how computational methods can advance sociological understanding of occupations as both structural and symbolic units of social life. Across three

projects, the dissertation has shown that the same analytic tools—embeddings, clustering, and large-scale text analysis—can illuminate distinct but interrelated aspects of work: its organization in the labor market, its representation in culture, and its role in social reproduction through partnership formation.

Structurally, the data-driven mapping of the Danish labor market demonstrates how occupations can be conceptualized not as fixed taxonomic entities but as dynamic positions in a network of mobility. This reframing bridges micro-level job transitions with macro-level class structure, revealing new patterns of stratification and resilience. Symbolically, the analysis of contemporary fiction uncovers how cultural hierarchies of work both reflect and distort real labor market inequalities, assigning visibility and desirability unevenly across professions. The exploratory analysis of assortative marriage links these two domains, showing how occupational proximity shapes intimate life and the reproduction of inequality.

Methodologically, the dissertation contributes to a growing body of work positioning computational techniques as theory-generating tools in sociology (Evans and Foster, 2019). It demonstrates how embeddings and LLM-assisted text analysis can serve as empirical instruments for mapping meaning, provided they are deployed with transparency, validation, and interpretive care. These methods extend sociological reach, enabling analyses that are both scalable and substantively grounded, and they exemplify how automation can complement, rather than supplant, theoretically-based social inquiry.

Taken together, the findings suggest that occupations occupy a dual life in modern societies: they organize economic opportunities and simultaneously structure the moral and cultural imagination of who we are and what kinds of work count as valuable. Computational methods make it possible to trace this duality empirically, bridging the material and the symbolic, the measurable and the meaningful. In doing so, this dissertation not only contributes to the sociology of work and culture but also points toward a more integrated, data-rich sociology—one capable of analyzing both the structures people inhabit and the stories they tell about them.

7 Conclusion

This dissertation set out to explore how computational methods can deepen understanding of occupations as both structural and symbolic foundations of social life. Across three interlinked projects, it examined how people move through work, how work is represented in culture, and how occupational proximity shapes patterns of partner matching. Each project addressed a different facet of the same underlying question: how occupations can be reimagined as dynamic, relational, and meaning-laden entities within an increasingly complex world of work.

Chapter 3 used population-level Danish registry data to derive an empirical map of the labor market based on observed career transitions. By embedding occupations in a vector space defined by real mobility patterns, it uncovered communities that reflect how people actually move between jobs rather than how experts classify them. This approach challenges static taxonomies and reveals a labor market structured less by formal categories than by flows of skill and opportunity. Chapter 4 turned to cultural data, applying large language model–assisted text analysis to more than a thousand bestselling novels to examine how fiction portrays work. The resulting picture showed striking symbolic hierarchies: certain occupations are narratively overrepresented while others are rendered invisible. The third, exploratory project used the embedding framework to revisit assortative marriage by occupation, demonstrating that continuous, data-driven measures of occupational distance provide a richer picture of homogamy than binary ”same-job” indicators.

Taken together, these studies show that the world of work can be analyzed through both its structures and its stories. Computational methods make it possible to trace this logic empirically, revealing the connective tissue between social structure and cultural imagination.

The dissertation’s broader contribution is methodological as well as substantive. It demonstrates how computational techniques can be used not simply to

process data but to build sociological theory from new forms of evidence. Embeddings, clustering, and hybrid human–machine coding each expand the empirical reach of sociology, but they also require reflexive engagement with questions of interpretation, bias, and ethics. The findings underscore that computational approaches are most powerful when paired with conceptual clarity and theoretical grounding: they allow us to see patterns, but it remains the researcher's task to explain why those patterns matter.

In closing, the projects presented here illustrate that understanding work in the twenty-first century requires attention not only to where people are employed, but to how occupations circulate through language, imagination, and desire. A future strand of sociological inquiry lies in connecting these dimensions: linking the data traces of everyday life to the narratives through which people make sense of them. Computational methods, when used critically, provide the means to do exactly that: to illuminate the deep patterns that structure both the realities and the representations of work.

Declarations

Funding

This work was funded in part (50%) by the University of Copenhagen, in part (25%) by a grant from Villum Fonden, and in part (25%) by a grant from the Novo Nordisk Foundation (grant number: NNF20OC0062897).

Conflicts of Interest

I declare that I have no conflicts of interest that could have influenced the research conducted or the results presented in this dissertation. All analyses, interpretations, and conclusions are my own and were carried out independently of any financial, personal, or professional relationships that could be construed as potential conflicts of interest.

Chapter 3 Supporting Information

A

A.1 word2vec Parameter Tuning

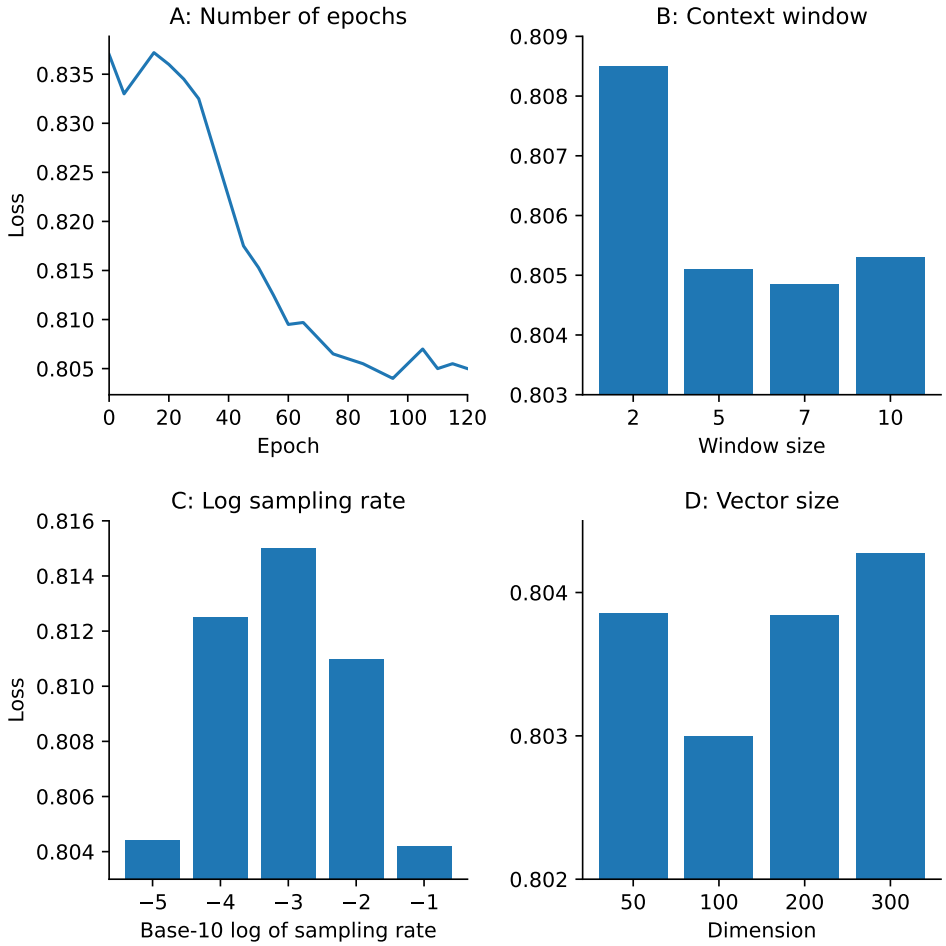

Figure A.1.. word2vec hyperparameter tuning based on validation loss. Model performance was evaluated using a predictive task: given a masked occupation code, the model predicted the missing token using learned embeddings. Each subplot shows validation loss across values of a single hyperparameter, with others held fixed. (A) Training epochs, (B) context window size, (C) base-10 logarithm of downsampling rate, (D) vector dimension.

A.2 Temporal Entropy, Disaggregated View

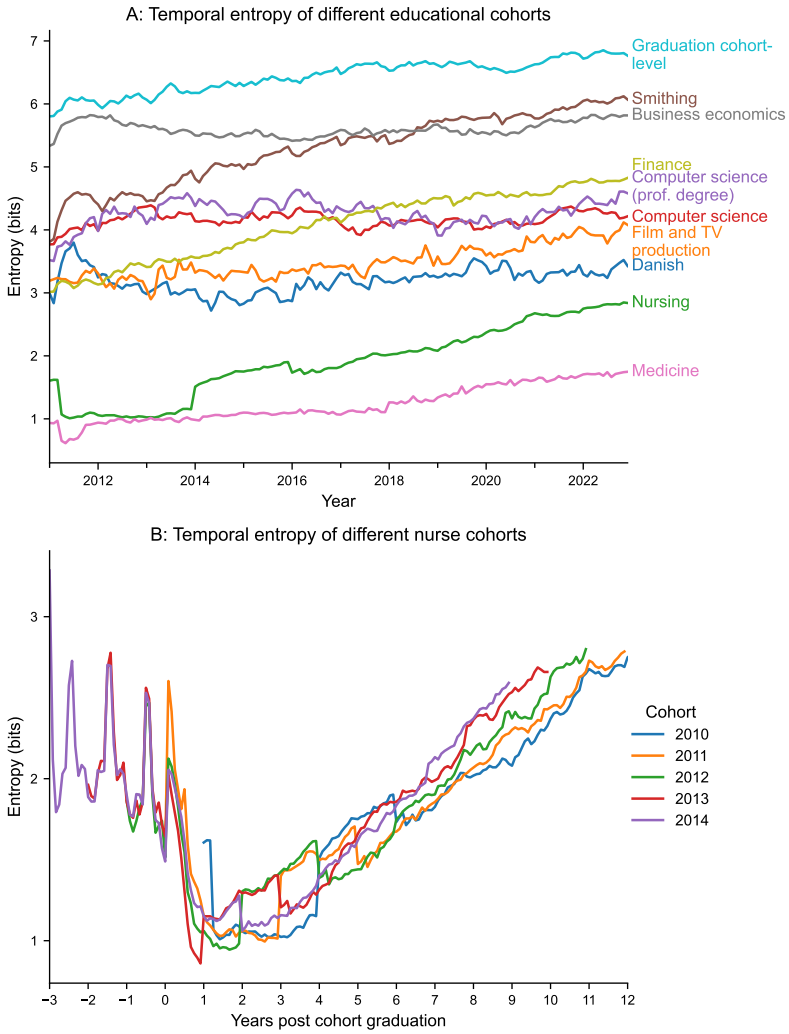

Figure A.2.. Temporal entropy of different educational cohorts, disaggregated into single occupations. (A) Temporal entropy for nine different educational cohorts who completed a degree in the indicated subject in 2010, plus a random sample of 2% of graduates in that year. (B) Temporal entropy for individuals who completed a nursing education in years 2010–2014 as a function of years post education. Both subplots depict entropy calculated over single occupation codes rather than occupation communities, as distinct from Fig 3.4.

A.3 Description of Data Files

All data files described below are available at the public repository `https://osf.io/hxrwt/overview?view_only=32f17694b5934d47882e0e236c2fa532`.

Community-level demographic and employment datasets. This archive (`S1_File.zip`) contains four CSV files providing additional datasets used in the analysis of Danish labor market communities (2011–2022):

- `node_community_mapping.csv` — Mapping of occupational codes (`NODE`) to their corresponding community identifiers (`COMMUNITY`), as derived from the occupational mobility clustering. Six-digit codes follow a Danish adaptation of ISCO (Statistics Denmark, 2011), where the ISCO code is given by the first four digits. Longer codes instead contain a NACE Rev. 2 code (Eurostat, 2008) followed by a suffix: `000110` indicates self-employment (including business ownership), `000120` indicates co-working spouses, and `000000` indicates employment in a firm with fewer than 10 employees. Community `-1` contains unclassified "noise" occupations. Occupations held by fewer than three individuals over the study period have been removed.

- `num_per_community.csv` — Monthly share of the total working population (ages 15–65) assigned to each community.

- `pct_male_per_community.csv` — Monthly percentage of male workers in each community.

- `pct_selfemp_per_community.csv` — Monthly percentage of workers in each community who are self-employed or registered as co-working spouses.

In the last three files, the first column (`MONTH`) gives the reference month, followed by one column per community (header = community identifier). Values are expressed in labor-hours rather than raw headcounts. To protect anonymity, any cells based on fewer than three individuals have been left blank.

Baseline age distribution of the labor force. This CSV file (`S2_File.parquet`) reports the aggregate age distribution of the Danish labor force (ages 15–65) across the full study period (2011–2022). Each row contains:

- `AGE` — worker age.
- `EXPECTED_SHARE` — share of total labor-hours contributed by workers of this age.

Values are expressed in labor-hours rather than headcounts, and provide the baseline reference distribution for community-level age analyses.

Community-level age composition. This Parquet file (`S3_File.parquet`) provides the age composition of each community over time (2011–2022), in long format. Each row corresponds to a community–month–age cell, with the following variables:

- `COMMUNITY` — community identifier.
- `MONTH` — reference month.
- `AGE` — worker age (15–65).
- `COMMUNITY_WEIGHT_AGE` — total labor-hours for workers of this age in the community.
- `COMMUNITY_TOTAL_HOURS` — total labor-hours for the community in that month.
- `PCT_WITHIN_COMMUNITY` — share of community labor-hours contributed by this age group.

To protect anonymity, any cell based on fewer than three individuals has been removed.

Median wage percentiles by age within communities. This Parquet file (`S4_File.parquet`) reports the median relative wage position of workers

by age within each community (2011–2022). Each row corresponds to a community–month–age cell, with the following variables:

- `COMMUNITY` — community identifier.
- `MONTH` — reference month.
- `AGE` — worker age (15–65).
- `MEDIAN_PERCENTILE` — median percentile rank of hourly wages within the community for this age group, calculated relative to other wage earners of the same age and month.

Cells with fewer than three contributing individuals have been removed to protect anonymity.

Community-level wage distributions. This Parquet file (`S5_File.parquet`) provides the wage distribution of each community over time (2011–2022), binned into fixed hourly wage intervals. Each row corresponds to a community–month–bin cell, with the following variables:

- `COMMUNITY` — community identifier.
- `MONTH` — reference month.
- `SALARY_BIN` — hourly wage bin in DKK (e.g. `[150,175)`). Values have not been inflation-adjusted or converted between currencies.
- `BIN_HOURS` — total labor-hours contributed by workers in this bin.
- `WAGE_TOTAL_HOURS` — total labor-hours of wage earners in the community that month.
- `PCT_WITHIN_WAGE_EARNERS` — share of community wage-earner labor-hours in this bin.

Values exclude self-employed and co-working spouses, for whom salary information is unavailable. Any cell based on fewer than three individuals (in the bin or its complement) has been removed for anonymity.

Chapter 4 Supporting Information

B

B.1 Prompting Strategy

I queried GPT-4o via the OpenAI API using a multi-stage prompting strategy combining model knowledge with external summaries. This process had three steps. First, the model was prompted to reason based on its own knowledge of the book, producing an initial output. Second, it was prompted again, this time with additional context from a Google search and any available blurbs or plot summaries. Finally, I combined results from both prompts, defaulting to the search-augmented output unless a field returned "Unknown," in which case I used the initial response. The prompt is displayed below, while the full script is available at `https://github.com/ella-clement/fiction-professions`.

```
You are a literary expert with deep knowledge of every
↪ best-selling book of the past decades.
Use your encyclopedic knowledge of books to retrieve detailed
↪ information.{search_context}{summary_context}{description_co
↪ ntext}

**Step 1: Determine the Genre and Main Setting**
- Identify the **genre** of the book. Provide a **one-word or
↪ short-phrase descriptor**.

**Step 2: Identify Protagonists (Only POV/Main Characters)**
- List only the **POV characters** or the **main narrators**.
- If the book features an **ensemble cast**, return **all
↪ protagonists**.
- Ensure that **supporting characters are not included**.

**Step 3: Assign Correct Professions (No Personality Traits)**
```

```
- Provide the **exact profession(s)** for each protagonist **in
 ↪  the same order**.
- If a character **changes professions**, list all relevant
 ↪  professions in **chronological order**.
- Always prefer **specific job titles** over general ones (e.g.,
 ↪  "Neurosurgeon" instead of "Doctor").
- **Strict rule**: Do not return personality-based words or
 ↪  identity markers that are not professions (e.g., "Kindest
 ↪  person" is invalid).
- If a profession is **completely unknown**, return "Unknown"
 ↪  instead of making assumptions.

**Step 4: Profession Mapping to ISCO Codes**
- Convert each profession into the **best ISCO code guess**.
- Ensure that ISCO codes follow the **same order as
 ↪  professions**.
- If a profession is unknown, return **0**. If unsure, return
 ↪  **9**.

**Step 5: Identify Love Interest (If Applicable)**
- Identify the **main love interest**, if applicable.
- Provide their **exact profession(s)** using the same rules as
 ↪  above.
- Convert their profession(s) into **ISCO codes**, following the
 ↪  same logic.

**Final Output Format (JSON):**
{{
  "Book Title": "{book_title}",
  "Book Author": "{book_author}",
  "Genre": "<genre>",
  "Protagonists": ["<name1>", "<name2>", ...],
  "Professions": [["<profession1a>", "<profession1b>"],
   ↪  ["<profession2a>"], ...],
  "ISCO": [["<isco1a>", "<isco1b>"], ["<isco2a>"], ...],
  "Love Interest": "<name or None>",
  "Love Interest Profession": ["<profession or None>"],
  "Love Interest's ISCO": ["<isco or None>"]
```

B.2 External Estimates for Rare Roles

For fictional roles not included in U.S. Bureau of Labor Statistics data (U.S. Bureau of Labor Statistics, 2024), I used approximate counts from secondary online sources. These estimates are order-of-magnitude only. I list the sources I used for each estimate below.

Table B.1.. External prevalence estimates for rare real-world roles in the U.S.

Fictional role	Source	Notes
Student	National Center for Education Statistics (n.d.)	
Secret agent	Federal Bureau of Investigation (2024) and PayScale (n.d.)	Exact estimate classified
Soldier	Statista (n.d.[a])	
Minor	United States Census Bureau (n.d.)	
Hospitality venue owner	Statista (n.d.[b]), market.us (2025), Statista (2025), FoodIndustry.Com (2022), and Glassdoor (2021b)	Estimate comes from chaining together multiple sources
Crime boss	Federal Bureau of Investigation (n.d.)	
Bookshop owner/worker	United States Census Bureau (2021) and Glassdoor (n.d.)	
Assassin	Adams (2013) and CDC (2025)	
Outside labor force	National Association of Home Builders (2025)	
Game designer	Clement (2025), Spira (2024), and Glassdoor (2021a)	Estimate comes from chaining together multiple sources
Personal assistant	Zippia (n.d.[b])	
Business owner	United States Census Bureau (2023)	

Continued on next page

Fictional role	Source	Notes
Navy/marine personnel	Tierney (2024) and indeed (n.d.)	
Sex worker	Sawicki *et al.* (2019)	
Bounty hunter	SkipNet Directory (2025)	Exact figure
Military leader	Statista (n.d.[a])	Exact estimate classified
Unemployed	National Association of Home Builders (2025)	
Prisoner	Carson and Kluckow (2023)	
Murderer	Statista (2024)	
Government employee	DeSilver (2025)	
Homemaker	National Association of Home Builders (2025)	
Content creator	MBO Partners (n.d.)	
Unspecified criminal	United States Sentencing Commission (2016)	
Camp counselor	Zippia (n.d.[a])	
Hotelier	United States Bureau of Labor Statistics (2025)	
Intern	Fennell (2022)	
Antique dealer	Better Business Bureau (n.d.)	
Astronaut	NASA (n.d.) and Dean (2018)	Exact figure
Occult/spiritual worker	Finley (2024)	
Hospital worker	Data USA (2017)	
Slave	Walk Free (n.d.)	
Home organizer	Professional Organizer Mavericks (2022)	
Hunter	ZipRecruiter (2025)	
Real estate developer	Baker Tilly (2023) and Glassdoor (2025b)	
Counterfeiter	United States Sentencing Commission (2014)	

Continued on next page

Fictional role	Source	Notes
Gig worker	United States Bureau of Labor Statistics (n.d.)	
Land owner	DeSilver (2021)	
Surrogate	Gonzalez (2019)	
Cryptocurrency miner	EUCI (2024)	
Motivational speaker	TCAA (2023)	
Religious leader	Luo (2006)	
Retiree	National Association of Home Builders (2025)	
Barista	Miniano (2025)	
Factory owner	IBISWorld (2025)	

C

Chapter 5 Supporting Information

C.1 Continuous Metric Comparison

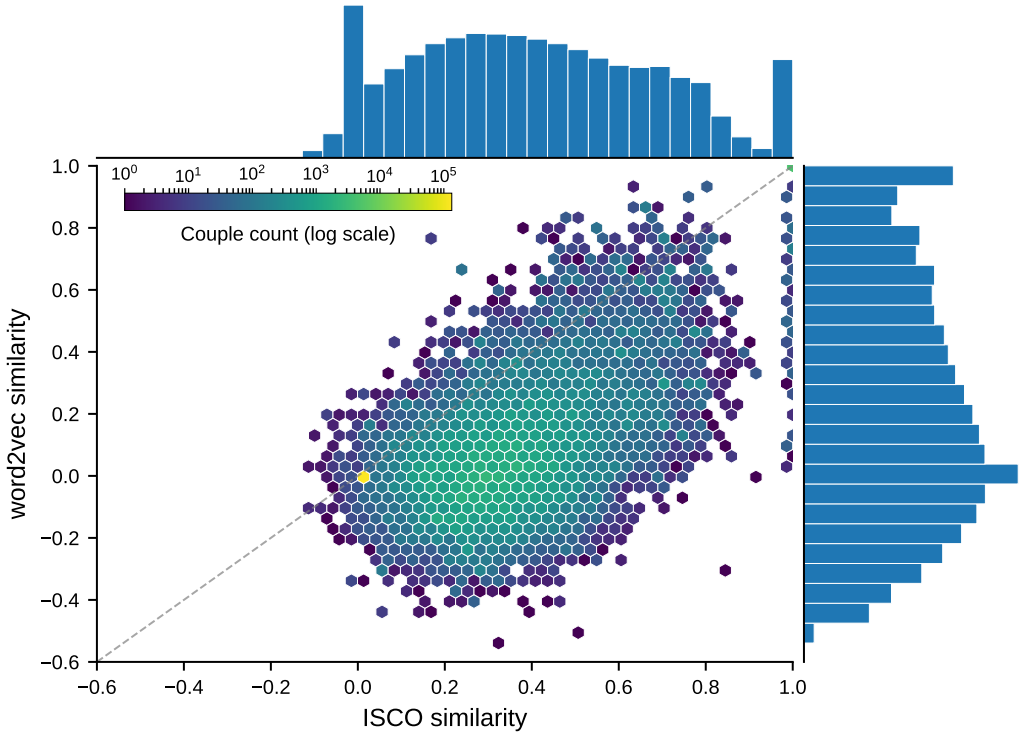

Figure C.1.. Joint distribution of ISCO and word2vec similarity measures at month of marriage. Hex-binned scatterplot of partners' occupation similarity at the month of marriage, with ISCO similarity on the horizontal axis and word2vec similarity on the vertical axis. Color intensity reflects the density of spouse pairs in each bin. Marginal histograms show the corresponding univariate distributions of the two metrics. Cells representing fewer than three observations have been removed to protect privacy.

C.2 Logistic Regression Tables

Table C.1: Logistic regression of marriage on indicator similarity.

Dep. Variable:	Y	**No. Observations:**	2089802
Model:	ConditionalLogit	**No. groups:**	189982
Method:	BFGS	**Min group size:**	11
Log-Likelihood:	-4.4295e+05	**Max group size:**	11
		Mean group size:	11.0

	coef	std err	z	P> |z|	[0.025	0.975]
x1	2.2223	0.013	173.791	0.000	2.197	2.247

Table C.3: Logistic regression of marriage on word2vec embedding similarity.

Dep. Variable:	Y	**No. Observations:**	2089802
Model:	ConditionalLogit	**No. groups:**	189982
Method:	BFGS	**Min group size:**	11
Log-Likelihood:	-4.3554e+05	**Max group size:**	11
		Mean group size:	11.0

	coef	std err	z	P> |z|	[0.025	0.975]
x1	0.2334	0.004	54.904	0.000	0.225	0.242
x2	0.2591	0.007	39.586	0.000	0.246	0.272

Table C.5: Logistic regression of marriage on word2vec and indicator similarity.

Dep. Variable:	Y	**No. Observations:**	2089802
Model:	ConditionalLogit	**No. groups:**	189982
Method:	BFGS	**Min group size:**	11
Log-Likelihood:	-4.3551e+05	**Max group size:**	11
		Mean group size:	11.0

	coef	std err	z	P> |z|	[0.025	0.975]
x1	-0.2486	0.032	-7.871	0.000	-0.311	-0.187

x2	0.2170	0.005	45.852	0.000	0.208	0.226
x3	0.3308	0.011	29.468	0.000	0.309	0.353

C.3 Description of Data Files

To facilitate reproducibility and secondary analysis, I provide three public Parquet files that contain aggregated measures of occupational similarity among married couples at the month of marriage. Each file corresponds to one of the three similarity metrics used in the analysis: exact occupational match ('indicator'), continuous similarity based on the ISCO hierarchy, and continuous similarity derived from the word2vec embedding model. All files here described are available at the following public repository: `https://osf.io/u8n3z/overview?view_only=e9be79f36c7e4cfc9c0bee23c8eb6f6c`.

Each file contains one row per combination of (i) calendar month of marriage and (ii) a binned similarity value for the given metric, together with the count of heterosexual couples married in that month who fall into that bin. Across all files, any couples whose occupational similarity is undefined (e.g., because one or both spouses are outside the labour force or have missing occupational codes) are grouped into a dedicated "missing" category.

- `assortative_similarity_indicator.parquet`. This file contains three columns:
  - `marriage_year_month`: a string of the form `YYYY-MM`,
  - `indicator_bin`: one of `"same"`, `"different"`, or `"missing"`,
  - `count`: the number of couples in that month–bin cell.
- `assortative_similarity_isco.parquet`. This file contains the same `marriage_year_month` and `count` columns, but replaces the indicator bin with a binned version of the continuous ISCO–based similarity score:
  - `isco_bin`: an interval label such as `[0.20,0.40)` or `[0.80,1.00]`, plus a `"missing"` category.

The bin edges are chosen to balance resolution with the need to avoid sparsity.

- `assortative_similarity_w2v.parquet`. This file is structured analogously, with:
  - `w2v_bin`: an interval label for the binned word2vec similarity score, plus a `"missing"` category.

Under the confidentiality rules of Statistics Denmark, it is not permitted to distribute individual-level microdata or aggregated outputs that contain cell counts below 3. These restrictions apply to all registries used in the analysis, including employment histories, occupation codes, and civil-status information. As a result, the fully disaggregated couple-level dataset used for the analyses in Chapter 5 cannot be released.

To comply with these requirements, the publicly released files contain only month-level aggregates and exclude any month–bin combination where the count of couples is strictly less than three. All similarity values are binned before aggregation to further reduce the risk of indirect identification.

Global References

Acemoglu, Daron and David Autor (2011). „Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings“. In: ed. by David Card and Orley Ashenfelter. Vol. 4. Handbook of Labor Economics. Elsevier, pp. 1043–1171.

Adams, Cecil (May 2013). „How Many People Get Killed for Money Each Year?“ In: *Washington City Paper*. Accessed: 2025-09-10.

Ahn, Yong-Yeol, James P. Bagrow, and Sune Lehmann (2010). „Link communities reveal multiscale complexity in networks“. In: *Nature*.

Andersen, Torben M. (2023). „The Danish labor market, 2000–2022“. In: *IZA World of Labor 2023* 404.

Baker Tilly (June 2023). *Capitalizing on Opportunities for Underrepresented Developers in Commercial Real Estate*. https://www.bakertilly.com/insights/capitalizing-on-opportunities-for-underrepresented-developers. Accessed: 2025-09-11.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). „Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk“. In: *Political Analysis* 20.3, pp. 351–368.

Better Business Bureau (n.d.). *Antique Dealers in USA*. https://www.bbb.org/. Accessed: 2025-09-11.

Bidwell, Matthew (Aug. 2013). „What Happened to Long-Term Employment? The Role of Worker Power and Environmental Turbulence in Explaining Declines in Worker Tenure“. In: *Organization Science* 24.

Binder, Matt (2016). *These Are the Most Popular Jobs on Tinder*. https://www.wired.com/story/tinder-popular-professions/. Accessed: 2025-04-22.

Blau, Peter M. and Otis Dudley Duncan (1967). *The American Occupational Structure*. New York: John Wiley and Sons.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001). „Chapter 59 - Measurement Error in Survey Data“. In: *Handbook of Econometrics*. Ed. by James J. Heckman and Edward Leamer. Vol. 5. Handbooks in Economics. Elsevier, pp. 3705–3843.

Cahuc, Pierre, Stéphane Carcillo, and André Zylberberg (2014). „Equilibrium Unemployment“. In: *Labor Economics*. 2nd. The MIT Press, pp. 553–574.

Carson, Anne and Rich Kluckow (Nov. 2023). *Prisoners in 2022 – Statistical Tables*. Tech. rep. 97. Bureau of Justice Statistics.

Cawelti, John G. (1976). *Adventure, Mystery, and Romance: Formula Stories as Art and Popular Culture*. Chicago: University of Chicago Press.

CDC (Jan. 2025). *Assault or Homicide*. `https://www.cdc.gov/nchs/fastats/homicide.htm`. Accessed: 2025-09-10.

Cheng, Siwei and Barum Park (Nov. 2020). „Flows and Boundaries: A Network Approach to Studying Occupational Mobility in the Labor Market“. In: *American Journal of Sociology* 126, pp. 577–631.

Chhetri, Tek Raj, Yibei Chen, Puja Trivedi, Dorota Jarecka, Saif Haobsh, Patrick Ray, Lydia Ng, and Satrajit S. Ghosh (2025). *STRUCTSENSE: A Task-Agnostic Agentic Framework for Structured Information Extraction with Human-In-The-Loop Evaluation and Benchmarking*. arXiv: `2507.03674`.

Clement, Jessica (Apr. 2025). *Topic: Video Game Industry in the United States*. `https://www.statista.com/topics/8790/video-game-industry-in-the-united-states/`. Accessed: 2025-09-10.

Coscia, Michele and Frank M. H. Neffke (2007). „Network Backboning with Noisy Data“. In: *IEEE 33rd International Conference on Data Engineering*, pp. 425–436.

Couldry, Nick (2012). *Media, Society, World: Social Theory and Digital Media Practice*. 1st. Cambridge: Polity.

Dahl, Christian Møller, Torben Johansen, and Christian Vedel (2024). *Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE*. arXiv: `2402.13604`.

Data USA (2017). *Hospitals*. `https://datausa.io/profile/naics/hospitals`. Accessed: 2025-09-11.

Dean, Brandi (Jan. 2018). „Becoming an Astronaut: Frequently Asked Questions“. In: *NASA*. Accesed: 2025-09-11.

DeSilver, Drew (Aug. 2021). „As National Eviction Ban Expires, a Look at Who Rents and Who Owns in the U.S.“ In: *Pew Research Center*. Accessed: 2025-09-11.

DeSilver, Drew (Jan. 2025). „What the Data Says about Federal Workers“. In: *Pew Research Center*. Accessed: 2025-09-11.

Dunkerley, David (1975). *Occupations and Society*. Routledge.

Education Connection (2022). *Sexiest Careers for Men and Women*. https://www.educationconnection.com/resources/sexiest-careers-for-men-and-women/. Accessed: 2025-04-22.

England, Paula (2005). „Emerging Theories of Care Work“. In: *Annual Review of Sociology* 31, pp. 381–399.

Erikson, Robert and John H. Goldthorpe (1993). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford University Press.

Erikson, Robert Clifford, John H Goldthorpe, and Lucienne Portocarero (1979). „Intergenerational Class Mobility in Three Western European Societies: England, France and Sweden“. In: *British Journal of Sociology* 30, p. 415.

Escobari, Marcela, Ian Seyal, and Carlos Baboin (June 2021). *Moving up: Promoting workers' upward mobility using network analysis*. Tech. rep. Brookings.

EUCI (Feb. 2024). *Facing a Growing Crypto-Mining Sector, the U.S. to Require Reports of Power Use by Miners*. https://www.euci.com/facing-a-growing-crypto-mining-sector-the-u-s-to-require-reports-of-power-use-by-miners/. Accessed: 2025-09-11.

EURES (EURopean Employment Services) (2025). *Labour Market Information: Denmark*. https://eures.europa.eu/living-and-working/labour-market-information/labour-market-information-denmark_en. Accessed: 2025-10-01.

Eurostat (2008). *NACE rev. 2: Statistical classification of economic activities in the European Community*. Luxembourg: European Commission.

Evans, James and Jacob G. Foster (2019). „Computation and the Sociological Imagination“. In: *Contexts* 18.4, pp. 10–15.

Federal Bureau of Investigation (2024). *Federal Bureau of Investigation Budget Request For Fiscal Year 2024*. https://www.fbi.gov/news/speeches-and-testimony/federal-bureau-of-investigation-budget-request-for-fiscal-year-2024. Testimony. Accessed: 2025-09-10.

Federal Bureau of Investigation (n.d.). *Gangs*. https://www.fbi.gov/investigate/violent-crime/gangs. Folder.

Fennell, Andrew (Dec. 2022). „Internship Statistics U.S.“ In: *StandOut CV*. Accessed: 2025-09-11.

Finley, Ben (June 2024). „Virginia City Repeals Ban on Psychic Readings as Industry Grows and Gains More Acceptance“. In: *AP News*. Accessed: 2025-09-11.

FoodIndustry.Com (Dec. 2022). *Despite the Pandemic, Independents Account for over 70% of All U.S. Restaurants*. https://www.foodindustry.com/articles/independents-account-for-70-percent-of-all-us-restaurants/. Accessed: 2025-09-10.

Ganzeboom, Harry B. G. (May 2010). „A New International Socio-Economic Index (ISEI) of Occupational Status for the International Standard Classification of Occupations 2008 (ISCO-08) Constructed with Data from the ISSP 2002–2007“. In: *Annual Conference of the International Social Survey Programme*. Lisbon, Portugal.

Ganzeboom, Harry B. G., Donald J. Treiman, and Wout C. Ultee (1991). „Comparative Intergenerational Stratification Research: Three Generations and Beyond“. In: *Annual Review of Sociology* 17, pp. 277–302.

Ganzeboom, Harry B.G. and Donald J. Treiman (1996). „Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations“. In: *Social Science Research* 25.3, pp. 201–239.

Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann (2018). „A network approach to topic models“. In: *Science Advances* 4.7, eaaq1360.

Glassdoor (2025a). https://www.glassdoor.com. Accessed: 2025-04-22.

Glassdoor (Aug. 2021a). *Salary: Game Designer in United States 2025*. https://www.glassdoor.com/Salaries/game-designer-salary-SRCH_KO0,13.htm. Accessed: 2025-09-10.

Glassdoor (Aug. 2021b). *Salary: Restaurant Owner in United States 2025*. https://www.glassdoor.com/Salaries/restaurant-owner-salary-SRCH_KO0,16.htm. Accessed: 2025-09-10.

Glassdoor (July 2025b). *Salary: Real Estate Developer in United States 2025*. https://www.glassdoor.com/Salaries/real-estate-developer-salary-SRCH_KO0,21.htm. Accessed: 2025-09-11.

Glassdoor (n.d.). *Salary: Bookstore Assistant in United States 2025*. https://www.glassdoor.com/Salaries/bookstore-assistant-salary-SRCH_KO0%2C19.htm. Accessed: 2025-09-10.

Glenn, Evelyn Nakano (1992). „From Servitude to Service Work: Historical Continuities in the Racial Division of Paid Reproductive Labor“. In: *Signs: Journal of Women in Culture and Society* 18.1, pp. 1–43.

Gonzalez, Alicia (2019). „Commercial Surrogacy in the United States“. In: *The Georgetown Journal of Gender and the Law* 21.1.

Gunz, Hugh, M.A. Peiperl, and Daniel Tzabbar (Jan. 2007). „Handbook of Career Studies“. In: SAGE Publications, Inc. Chap. Boundaries in the study of career, pp. 471–494.

Holtzman, Linda and Leon Sharpe (2014). *Media Messages: What Film, Television, and Popular Music Teach Us About Race, Class, Gender, and Sexual Orientation*. 2nd. New York: Routledge, p. 558.

Hughes, Bradley T., Sanjay Srivastava, Magdalena Leszko, and David M. Condon (Feb. 2024). „Occupational Prestige: The Status Component of Socioeconomic Status“. In: *Collabra: Psychology* 10.1, p. 92882.

IBISWorld (Aug. 2025). *Manufacturing in the US: Number of Businesses Statistics*. https://www.ibisworld.com/united-states/number-of-businesses/manufacturing/210/. Accessed: 2025-09-11.

indeed (n.d.). *Petty Officer Salaries in the United States for US Navy*. https://www.indeed.com/cmp/US-Navy-faba947a/salaries/Petty-Officer. Accessed: 2025-09-11.

International Labour Office (2023). *The International Standard Classification of Occupations (ISCO-08) Companion Guide*. Geneva: International Labour Office.

Kalleberg, Arne L. (2009). „Precarious Work, Insecure Workers: Employment Relations in Transition“. In: *American Sociological Review* 74.1, pp. 1–22.

Kalmijn, Matthijs (1994). „Assortative Mating by Cultural and Economic Occupational Status“. In: *American Journal of Sociology* 100.2, pp. 422–452.

Kalmijn, Matthijs (1998). „Intermarriage and Homogamy: Causes, Patterns, Trends“. In: *Annual Review of Sociology* 24.Volume 24, 1998, pp. 395–421.

Kim, Tae-Yeon, Seong-Uk Baek, Myeong-Hun Lim, *et al.* (2024). „Occupation classification model based on DistilKoBERT: using the 5th and 6th Korean Working Condition Surveys“. In: *Annals of Occupational and Environmental Medicine* 36, e19.

Kozlowski, Austin C., Matt Taddy, and James A. Evans (Sept. 2019). „The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings“. In: *American Sociological Review* 84.5, pp. 905–949.

Larson, Magali Sarfatti (1977). *The Rise of Professionalism: A Sociological Analysis*. 1st ed. University of California Press.

Luo, Michael (Aug. 2006). „Seeking Entry-Level Prophet: Burning Bush and Tablets Not Required“. In: *The New York Times*. Accessed: 2025-09-11.

Luo, S. (2017). „Assortative mating and couple similarity: Patterns, mechanisms, and consequences“. In: *Social and Personality Psychology Compass* 11.8, e12337.

Macanovic, Ana (2022). „Text mining for social science: The state and the future of computational text analysis in sociology“. In: *Social Science Research* 108, p. 102784.

Mäkelä, Elina and Fabian Stephany (2025). *Complement or substitute? How AI increases the demand for human skills*. arXiv: 2412.19754.

market.us (Feb. 2025). *Pubs, Bars and Nightclubs Market*. Tech. rep. 139076. Accessed: 2025-09-10. market.us, p. 364.

Mazumder, Bhashkar and Miguel Acosta (2015). „Using Occupation to Measure Intergenerational Mobility“. In: *The Annals of the American Academy of Political and Social Science* 657, pp. 174–193.

MBO Partners (n.d.). *Creator Economy Trends Report 2024*. https://www.mbopartners.com/state-of-independence/creator-economy-report/. Accessed: 2025-09-11.

McInnes, Leland, John Healy, and Steve Astels (Mar. 2017). „hdbscan: Hierarchical density based clustering“. In: *Journal of Open Source Software* 2.11, p. 205.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (2001). „Birds of a Feather: Homophily in Social Networks“. In: *Annual Review of Sociology* 27, pp. 415–444.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781.

Miniano, Marcy (July 2025). „Barista Statistics in 2025: Salaries, Career Stats, Consumer Trends & Industry Growth“. In: *OysterLink*. Accessed: 2025-09-11.

Morris, Kathy (2021). *The Most (And Least) Attractive Jobs for a Romantic Partner*. https://www.zippia.com/advice/most-attractive-jobs-least/. Accessed: 2025-04-22.

NASA (n.d.). *Active Astronauts*. https://www.nasa.gov/humans-in-space/astronauts/active-astronauts/. Accessed: 2025-09-11.

National Association of Home Builders (Mar. 2025). *People Not in the Labor Force*. https://eyeonhousing.org/2025/03/people-not-in-the-labor-force/. Accessed: 2025-09-10.

National Center for Education Statistics (n.d.). *Fast Facts*. https://nces.ed.gov/fastfacts/display.asp?id=372. Accessed: 2025-09-10.

Neuendorf, Kimberly A. (2017). „Reliability“. In: *The Content Analysis Guidebook*. 2nd ed. Chapter 6. SAGE Publications, Inc.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, *et al.* (2020). „Bias in data-driven artificial intelligence systems—An introductory survey“. In: *WIREs Data Mining and Knowledge Discovery* 10.3.

Nwosu, Chijioke O, Umakrishnan Kollamparambil, and Adeola Oyenubi (2022). „Socio-economic inequalities in ability to work from home during the coronavirus pandemic“. In: *The Economic and Labour Relations Review* 33.2, pp. 290–307.

Oesch, Daniel (2006). „Coming to Grips with a Changing Class Structure: An Analysis of Employment Stratification in Britain, Germany, Sweden and Switzerland“. In: *International Sociology* 21.2, pp. 263–288.

OpenAI (2024). *GPT-4o System Card*. Tech. rep. OpenAI.

PayScale (n.d.). *FBI Agent Salary in 2025*. https://www.payscale.com/research/US/Job=FBI_Agent/Salary. Accessed: 2025-09-10.

Pedersen, Carsten Bøcker (2011). „The Danish Civil Registration System“. In: *Scandinavian Journal of Public Health* 39.7_suppl, pp. 22–25.

Persson, Christian, Ludvig Bohlin, Daniel Edler, and Martin Rosvall (2016). *Maps of sparse Markov chains efficiently reveal community structure in network flows with memory*. arXiv: 1606.08328.

Professional Organizer Mavericks (Aug. 2022). *Is Professional Organizing A Lucrative Business?* https://professionalorganizermavericks.com/start/is-professional-organizing-a-lucrative-business/. Accessed: 2025-09-11.

Reimers, Nils and Iryna Gurevych (Nov. 2019). „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks“. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rio-Chanona, R. Maria del, Penny Mealy, Mariano Beguerisse-Díaz, François Lafond, and J. Doyne Farmer (2020). „Occupational mobility and automation: a data-driven network model“. In: *Journal of the Royal Society Interface*.

Rosvall, Martin and Carl T. Bergstrom (2008). „Maps of Random Walks on Complex Networks Reveal Community Structure“. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.4, pp. 1118–1123.

Savcisens, Germans, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann (2024). „Using sequences of life-events to predict human lives“. In: *Nature Computational Science* 4.1, pp. 43–56.

Sawicki, Danielle A., Brienna N. Meffert, Kate Read, and Adrienne J. Heinz (Feb. 2019). „Culturally Competent Health Care for Sex Workers: An Examination of Myths That Stigmatize Sex-Work and Hinder Access to Care“. In: *Sexual*

and relationship therapy : journal of the British Association for Sexual and Relationship Therapy 34.3, pp. 355–371.

Schöch, Christof (2021). *Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama*. arXiv: 2103.13019.

Schwartz, Christine and Nikki Graf (2009). „Assortative matching among same-sex and different-sex couples in the United States, 1990-2000“. In: *Demographic Research* 21.28, pp. 843–878.

Schwartz, Christine, Yu Wang, and Robert Mare (Aug. 2021). „Opportunity and change in occupational assortative mating“. In: *Social Science Research* 99, p. 102600.

Serrano, M. Ángeles, Marián Boguñá, and Alessandro Vespignani (2009). „Extracting the multiscale backbone of complex weighted networks“. In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6483–6488.

Siddiqui, Nabeel (Jan. 2024). „Cutting the Frame: An In-Depth Look at the Hitchcock Computer Vision Dataset“. In: *Journal of Open Humanities Data* 10.

Simhi, Adi and Shaul Markovitch (2023). *Interpreting Embedding Spaces by Conceptualization*. arXiv: 2209.00445.

SkipNet Directory (2025). *SkipNet Directory*. https://nafra.info/. Accessed: 2025-09-11.

Sørensen, Jesper B. and David B. Grusky (1996). „The Structure of Career Mobility in Microscopic Perspective“. In: *Social Differentiation And Social Inequality*. Routledge.

Sørensen, Susanne Mainz (2024). *Statistikdokumentation for Elevregistret 2024*. Tech. rep. Statistics Denmark.

Spira, Lisa (Dec. 2024). *US Escape Room Industry Report – December 2024*. Tech. rep. Accessed: 2025-09-10. Room Escape Artist.

Statista (Sept. 2024). *Murder in the U.S.: Number of Offenders by Gender 2023*. https://www.statista.com/statistics/251886/murder-offenders-in-the-us-by-gender/. Accessed: 2025-09-11.

Statista (Feb. 2025). *Topic: Coffee Shops and Cafes in the U.S.* https://www.statista.com/topics/1670/coffeehouse-chain-market/. Accessed: 2025-09-10.

Statista (n.d.[a]). *Army Military Personnel by Rank U.S. 2025*. https://www.statista.com/statistics/239383/total-military-personnel-of-the-us-army-by-grade/. Accessed: 2025-09-10.

Statista (n.d.[b]). *Number of Coffee Shops US 2022*. https://www.statista.com/statistics/1000058/number-of-coffeehouse-stores-in-the-us/. Accessed: 2025-09-10.

Statistics Denmark (Mar. 2011). *DISCO-08: Danmarks Statistiks fagklassifikation*. Technical manual. First edition. Statistics Denmark.

Statistics Denmark (2025). *Population Figures*. https://www.dst.dk/en/Statistik/emner/borgere/befolkning/befolkningstal. Accessed: 2025-05-05.

Stender, Pernille, Thomas Thorsen, and Hans Henrik Andersen (2015). „Micro data integration for Labour Market Account". In: *Statistical Journal of the IAOS*.

Steyvers, Mark and Joshua B. Tenenbaum (2001). *The large-scale structure of semantic networks: statistical analyses and a model for semantic growth*. arXiv: cond-mat/0110012.

TCAA (Oct. 2023). *Motivational Speakers in USA: Top 10 Revealed for 2024!* https://www.tcaa.co/motivational-speakers-in-usa/. Accessed: 2025-09-11.

Thrun, M. C. and A. Ultsch (2021). „Using Projection-Based Clustering to Find Distance- and Density-Based Clusters in High-Dimensional Data". In: *J Classif* 38, pp. 280–312.

Tierney, Abigail (Nov. 2024). *Military Force Numbers by Service Branch and Reserve Component U.S. 2023*. https://www.statista.com/statistics/232330/us-military-force-numbers-by-service-branch-and-reserve-component/. Accessed: 2025-09-11.

Toubøl, Jonas and Anton Grau Larsen (2017). „Mapping the Social Class Structure: From Occupational Mobility to Social Class Categories Using Network Analysis". In: *Sociology* 51.6, pp. 1257–1276.

Treiman, Donald J. (1977). *Occupational Prestige in Comparative Perspective*. English. Quantitative studies in social relations. New York: Academic Press.

U.S. Bureau of Labor Statistics (2024). *Occupational Employment and Wage Statistics by Industry*. https://data.bls.gov/oes/#/industry/000000. Accessed: 2025-04-22.

United States Bureau of Labor Statistics (Aug. 2025). *Lodging Managers*. https://www.bls.gov/ooh/management/lodging-managers.htm. Accessed: 2025-09-11.

United States Bureau of Labor Statistics (n.d.). *Contingent and Alternative Employment Arrangements Summary*. https://www.bls.gov/news.release/conemp.nr0.htm. Accessed: 2025-09-11.

United States Census Bureau (Dec. 2021). *Don't Turn the Page on Bookstores*. https://www.census.gov/library/stories/2021/12/do-not-turn-the-page-on-bookstores.html. Accessed: 2025-09-10.

United States Census Bureau (Oct. 2023). *Census Bureau Releases New Data on Minority-Owned, Veteran-Owned and Women-Owned Businesses*. https://www.census.gov/newsroom/press-releases/2023/annual-business-survey-employer-business-characteristics.html. Accessed: 2025-09-10.

United States Census Bureau (n.d.). *U.S. Census Bureau QuickFacts*. https://www.census.gov/quickfacts/fact/table/US/PST045224. Accessed: 2025-09-10.

United States Sentencing Commission (2014). *Quick Facts: Counterfeiting Offenses*. Tech. rep. United States Sentencing Commission.

United States Sentencing Commission (Mar. 2016). *Career Offenders*. https://www.ussc.gov/research/quick-facts/career-offenders. Accessed: 2025-09-11.

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei (2024). *CAREER: A Foundation Model for Labor Sequence Data*. arXiv: 2202.08370.

Villarreal, Andrés (2020). „The U.S. Occupational Structure: A Social Network Approach". In: *Sociological Science* 7.8, pp. 187–221.

Walk Free (n.d.). *Modern Slavery in United States*. https://www.walkfree.org/global-slavery-index/country-studies/united-states/. Accessed: 2025-09-11.

Zimmermann, Mikkel (2025). *Unemployed Persons*. https://www.dst.dk/en/Statistik/emner/arbejde-og-indkomst/beskaeftigelse-og-arbejdsloeshed/arbejdsloese. Accessed: 2025-10-01.

Zippia (2025). https://https://www.zippia.com/. Accessed: 2025-04-22.

Zippia (n.d.[a]). *Camp Counselor Demographics and Statistics in the US*. https://www.zippia.com/camp-counselor-jobs/demographics/. Accessed: 2025-09-11.

Zippia (n.d.[b]). *Job outlook for personal assistants in the United States*. https://www.zippia.com/personal-assistant-jobs/trends/. Accessed: 2025-09-12.

ZipRecruiter (Sept. 2025). *Salary: Professional Hunter*. https://www.ziprecruiter.com/Salaries/Professional-Hunter-Salary. Accessed: 2025-09-11.