# What is a Labor Market?
# Classifying Workers and Jobs Using Network Theory

by

James Skidmore Fogel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2022

Doctoral Committee:

Professor Matthew D. Shapiro, Chair
Professor John Bound
Professor Abigail Jacobs
Professor Sebastian Sotelo

James Skidmore Fogel
jsfog@umich.edu
ORCID ID 0000-0003-0051-504X

# ACKNOWLEDGEMENTS

I would first like to thank my dissertation committee. My chair, Matthew Shapiro, believed in my unconventional idea from the start, well before I could articulate it coherently, and provided me with invaluable guidance and encouragement throughout the entire process. My entire committee — Matthew Shapiro, Abigail Jacobs, John Bound, and Sebastian Sotelo — always believed in me and helped me shape my vision into a reality. I am indebted to the countless other professors, classmates, seminar participants, and friends who pushed me, taught me, and supported me throughout my PhD.

I want to thank my family and friends who kept me not just sane, but happy. One of the main reasons I came to the University of Michigan for graduate school was that I knew it would be a place where I would be happy, surrounded by family and friends. Seven years later I feel totally justified in that decision. To my family, the friends I already had, and the new friends I made along the way, I treasure all of the time I have spent and will continue to spend with you, and I couldn't have done it without you.

Finally, I want to thank my friend and co-author Bernardo Modenesi. We collaborated on nearly every aspect of this dissertation and I absolutely could not have done it without him. He has always been patient, thoughtful, intelligent, generous, and kind, not to mention a pleasure to work with. There are many problems I could not have solved without Bernardo's help, and perhaps more importantly, there were many frustrations and low points that I could not have navigated without Bernardo's friendship and support.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1256260. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research is also supported by the Alfred P. Sloan Foundation through the CenHRS project at the University of Michigan. Chapters I and II were done in partnership with the Brazilian Institute of Applied Economic Research (IPEA). Rafael Pereira at IPEA provided invaluable support throughout this process.

Chapter III uses data from the U.S. Census Bureau. Any views expressed are

those of the authors and not those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this information product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2176. (CBDRB-FY22-P2176-R9773). This research uses data from the Census Bureau's Longitudinal Employer Household Dynamics Program, which was partially supported by the following National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

# TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

**Table**

# LIST OF APPENDICES

**Appendix**

# ABSTRACT

This dissertation combines economic theory and network theory to develop a new methodology for identifying latent worker and job heterogeneity from the network of worker–job matches in linked employer-employee data sets. Chapter I develops most of the theory and describes the methodology in detail before applying it to estimating the effects of labor demand shocks on workers. Chapter II extends the methodology developed in Chapter I and applies it to gender wage gap decompositions. Finally, Chapter III uses the methodology from Chapter I to impute occupation on the U.S. Census Bureau's Longitudinal Employer Household Dynamics (LEHD) data set.

Chapter I, which is co-authored with Bernardo Modenesi, develops a new data-driven approach to characterizing latent worker skill and job task heterogeneity by applying an empirical tool from network theory to large-scale Brazilian administrative data on worker–job matching. It microfounds this tool using a standard equilibrium model of workers matching with jobs according to comparative advantage. The classifications identify important dimensions of worker and job heterogeneity that standard classifications based on occupations and sectors miss. The equilibrium model based on these classifications more accurately predicts wage changes in response to the 2016 Olympics than a model based on occupations and sectors. Additionally, for a large simulated shock to demand for workers, the chapter shows that reduced form estimates of the effects of labor market shock exposure on workers' earnings are nearly 4 times larger when workers and jobs are classified using these new classifications as opposed to occupations and sectors.

Chapter II, which is co-authored with Bernardo Modenesi, measures gender discrimination by decomposing male–female differences in average wages into a component explained by male and female workers having different productivity distributions and a component explained by equally productive male and female workers being paid differently. This requires researchers to build reliable counterfactuals by identifying all relevant controls such that male workers are compared to female workers who are identical in all aspects relevant to pay other than their gender, conditional on controls. To do this, this chapter (i) develops a new economically principled network-based approach to control for unobserved worker skill and job task heterogeneity using the information revealed by detailed data

on worker–job matching patterns, (ii) non-parametrically estimates counterfactual wage functions for male and female workers, (iii) introduces a correction for the possibility that the male and female productivity distributions do not overlap, and (iv) applies these new methods by revisiting gender wage gap decompositions using improved counterfactuals based on (i), (ii) and (iii). The chapter decomposes the gender wage gap in Rio de Janeiro, Brazil and finds that the gender wage gap is almost entirely explained by male and female workers who possess similar skills and perform similar tasks being paid different wages.

Chapter III attempts to impute occupation on the LEHD by exploiting the information contained in the LEHD's rich set of worker–job matches using the method developed in Chapter I. It finds that while the information contained in these matches is informative about economic outcomes like earnings, it is minimally informative about occupation. In particular, the information gleaned from worker–job matches has minimal predictive power for occupation when other variables like industry are included as predictors.

# CHAPTER I

# What is a Labor Market? Classifying Workers and Jobs Using Network Theory (with Bernardo Modenesi)

## 1.1 Introduction

Many questions in economics lead researchers to classify heterogeneous workers and jobs into discrete groups. For example, to estimate the effect of a labor supply or demand shock on workers, researchers identify groups of similar workers who they assume to have had the same exposure to the shock and compare outcomes between differentially exposed groups of workers.[1] Similarly, to characterize two-sided (worker–job) multidimensional heterogeneity, researchers identify groups of workers with similar skills and study how they match with groups of jobs requiring similar tasks.[2] Studies of labor market power compute the concentration of individual firms within groups of similar jobs that compete with each other for labor.[3] The standard approach to characterizing heterogeneity is to group workers and/or jobs based on observable variables such as age, education, occupation, industry, or geography. This approach has limitations: (i) relevant dimensions of worker and job heterogeneity may be unobserved or measured with error, and (ii) it requires researchers to decide which dimensions of heterogeneity are important.[4] This paper proposes a new model-consistent and data-driven approach to characterizing worker and

[1] For example, Autor et al. (2013) group workers by commuting zone, and Card (1990) groups workers by race and predicted earnings quartile.

[2] Autor et al. (2003); Acemoglu and Autor (2011); Autor (2013); Tan (2018); Lindenlaub (2017); Kantenga (2018)

[3] Azar et al. (2018); Benmelech et al. (2018); Rinz (2018); Azar et al. (2019); Schubert et al. (2020); Arnold (2020); Lipsius (2018); Jarosch et al. (2019)

[4] A related approach uses direct measures of skills and tasks from sources such as the Occupational Information Network (O*NET) or Dictionary of Occupational Titles (DOT). For a discussion of the limitations of this approach, see Frank et al. (2019) who note that "according to O*NET, the skill 'installation' is equally important to both computer programmers and to plumbers, but, undoubtedly, workers in these occupations are performing very dissimilar tasks."

job heterogeneity. In an empirical application it demonstrates that using traditional worker and job classifications in Bartik-style regressions leads us to significantly understate the effect of exposure to shocks on workers' earnings.

We employ a revealed preference approach that relies on workers' and jobs' choices, rather than observable variables or expert judgments, to classify workers and jobs. Our key insight is that linked employer-employee data contain a previously underutilized source of information: millions of worker–job matches, each of which reflects workers' and jobs' perceptions of the workers' skills and the jobs' tasks. Intuitively, if two workers are employed by the same job, they probably have similar skills, and if two jobs employ the same worker those jobs probably require workers to perform similar tasks.

We formalize this intuition and apply it to large-scale data using a Roy (1951) model in which workers supply labor to jobs according to comparative advantage. Workers belong to a discrete set of latent *worker types* defined by having the same "skills" and jobs belong to a discrete set of latent *markets* defined by requiring employees to perform the same "tasks."[5] Workers match with jobs according to comparative advantage, which is determined by complementarities between skills and tasks at the worker type–market level. Workers who have similar vectors of match probabilities over markets are therefore revealed to have similar skills and belong to the same worker type, and jobs that have similar vectors of match probabilities over worker types are revealed to have similar tasks and belong to the same market.

In an ideal data set we would observe each worker choosing jobs an infinite number of times, allowing us to observe the exact worker–job match probability distribution. Since this is infeasible, we use a tool from the community detection branch of network theory called the bipartite stochastic block model (BiSBM). The BiSBM uses realized job matches of each worker's peers — coworkers, former coworkers, coworkers' former coworkers, former coworkers' coworkers, and so on — as proxies for that worker's match probability distribution over jobs, and uses these match probabilities to classify workers and jobs into worker types and markets. Our model microfounds the BiSBM, giving the worker types and markets it identifies a rigorous theoretical underpinning and clear interpretability.

Once we have assigned workers to worker types and jobs to markets, we estimate the parameters of the labor supply Roy model and embed it in a general equilibrium model with workers, firms, households and exogenous product demand shocks, which propagate through the model to generate labor demand shocks. The key parameter of the model is a matrix defining the productivity of each worker type when employed in each market. We estimate

[5] "Skills" and "tasks" should be interpreted broadly as any worker and job characteristics that determine which workers match with which jobs.

the productivity matrix using a maximum likelihood procedure that formalizes the intuition that worker type–market matches that (i) occur more frequently and (ii) pay higher wages are revealed to be more productive.

We estimate our model and conduct empirical analyses using Brazilian adminstrative records from the Annual Social Information Survey (RAIS) that is managed by the Brazilian labor ministry. The RAIS data contain detailed information about every formal sector employment contract, including worker demographic information, occupation, sector, and earnings. Critically, these data represent a network of worker–job matches in which workers are connected to every job they have ever held, allowing us to identify job histories of workers, their coworkers, their coworkers' coworkers, and so on. We restrict our analysis to the Rio de Janeiro metropolitan area both for computational reasons and because restricting to a single metropolitan area enables us to focus on skills and tasks dimensions of worker and job heterogeneity rather than geographic heterogeneity. While many others have used linked employer-employee data (LEED), we are the first to fully utilize the rich information embedded in the network of worker–job matches.[6]

Our novel approach to characterizing fine-grained worker and job heterogeneity revealed by LEED allows us to reevaluate the effects of labor market shocks on workers and consider how sensitive results are to the way workers and jobs are classified. We do this using both structural and reduced form methods. In the structural approach, we use our general equilibrium model to simulate the effect of the 2016 Rio de Janeiro Olympics on workers' earnings. We show that a model based on worker types and markets more accurately predicts actual Olympics-induced changes in workers' earnings than a series of benchmarks in which we use the same model but define worker and job heterogeneity using more traditional approaches based on occupation and sector.

Next, we apply our classifications to reduced form Bartik-style regressions and find that our method significantly increases estimates of the effects of workers' exposure to labor market shocks on their earnings. We estimate the effect of the 2016 Olympics on workers and show that both coefficient estimates and $R^2$ values are significantly larger when workers and jobs are classified using our worker types and markets as opposed to occupations and sectors. We then perform a series of simulations in which we feed shocks through our model to generate data in which we know the true data generating process and estimate the effects of the shocks on workers in the simulated data using our network-based classifications and using conventional classifications. Across these simulations, the estimated effects of the shocks on workers' earnings are on average 3.7 times larger using our classifications as opposed to

[6]Nimczik (2018) and Jarosch et al. (2019) use a related method to classify firms using a unipartite network of firms linked by worker transitions, however they do not classify individual workers or jobs.

conventional classifications. Finally, we perform a detailed case study of a simulated shock to understand why our classifications outperform traditional ones. We show that our worker types more precisely identify groups of workers who experienced similar exposure to labor market shocks than do occupations and our markets more precisely identify groups of jobs that hire similar workers than do sectors.

In a series of descriptive analyses, we provide supporting evidence that helps explain why conventional methods may understate the effects of shocks on workers. We show that our worker types and markets capture meaningful information about the worker and job characteristics relevant for labor market outcomes that conventional classifications miss. First, we demonstrate that our worker types aggregate workers across distinct occupations who are revealed to have similar skills, while simultaneously disaggregating workers in the same occupation with different skills. For example, we find that coaches and physical education teachers belong to the same worker type, while physical education and math teachers do not. Second, we show that our worker types do a better job of maximizing within-group skill homogeneity and between-group skill heterogeneity than do 4-digit occupations. Third, we show that worker types' labor supply is more concentrated within markets than within sectors, indicating that markets outperform sectors in terms of identifying groups of jobs that are similar from the perspective of workers.

**Literature:** We contribute to the large literature measuring the effects of labor market shocks on workers using either reduced form methods (Autor et al., 2013; Card, 1990; Autor et al., 2014; Yagan, 2017; Bound and Holzer, 2000; Blanchard and Katz, 1992; Bartik, 1991), or a structural approach (Burstein et al., 2019; Caliendo et al., 2019; Galle et al., 2017; Kim and Vogel, 2021). Relative to both of these literatures, our contribution is a new approach to classifying workers and jobs based on latent heterogeneity.

Conditional on assigning workers to latent worker types and jobs to latent markets, our model of labor supply is similar to Grigsby (2019) and Bonhomme et al. (2019). Our method for clustering workers and jobs builds upon the bipartite stochastic block model from the community detection branch of the network theory literature (Larremore et al., 2014; Peixoto, 2019). A major contribution of our paper is creating a theoretical link between a labor supply model and the BiSBM, thereby providing microfoundations for using tools from network theory to solve problems in economics and giving these tools clear economic interpretability.

Like Sorkin (2018), Nimczik (2018), and Jarosch et al. (2019), we use tools from network theory to extract previously unobserved information from LEED. We use the panel of worker–job matches to identify worker and job *similarities*; by contrast, Sorkin exploits the direction of worker flows between firms to identify *differences* between firms. Nimczik

(2018), and Jarosch et al. (2019) are also interested in using network data to identify similarities, however they cluster together only firms, abstracting from worker heterogeneity and within-firm job heterogeneity, while we cluster workers *and* jobs simultaneously. Schmutte (2014) uses a different tool from network theory to cluster workers and firms using survey data, however our microfoundations and detailed data allow us to identify more fine-grained heterogeneity and provide model-based interpretability of our classifications.

Our approach to modeling multidimensional worker–job heterogeneity is related to the literature on worker–job matching in a skills-tasks framework (Autor et al., 2003; Acemoglu and Autor, 2011; Autor, 2013; Lindenlaub, 2017; Tan, 2018; Kantenga, 2018). Relative to this literature, we provide a theoretically principled and data-driven way of identifying groups of workers with similar skills and groups of jobs with similar tasks. Mansfield (2019) also studies two-sided matching and integrates skill–task dimensions with geographic dimensions. Our contribution is to improve identification of clusters of workers and jobs who are similar in terms of high-dimensional latent skills and tasks, respectively.

**Roadmap:** The paper proceeds as follows. Section 1.2 lays out our economic model. Section 1.3 builds upon the model to derive a maximum likelihood procedure for clustering workers into worker types and jobs into markets. Section 1.4 derives a maximum likelihood estimator for labor supply parameters, including a matrix of worker type–market match productivities. Section 1.5 discusses our data and sample restrictions. Section 1.6 presents summary statistics from our worker and job classification method. Section 1.7 shows that a version of our equilibrium model based on our network-based worker and job classifications is better at predicting the effects of a real world shock than one based on standard classifications. Section 1.8 applies our classifications to Bartik-style regressions and shows that standard methods may be understating the effects of shocks on workers. Section 1.9 concludes.

## 1.2 Model

In this section we develop a model that is suited to analyzing data containing high resolution information on worker–job matches. We describe our data in detail in Section 1.5.

### 1.2.1 Model set up

We propose a model with three primary components: heterogeneous workers who supply labor, heterogeneous sectors each composed of competitive firms producing a sector-specific good, and a representative household which consumes firms' output. Workers supply their skills to jobs, which are bundles of tasks. Jobs' tasks are combined by the firms' production

functions to produce output. The most important part of the model is the labor market, which has the following components:

- Each worker is endowed with a *worker type*, and all workers of the same type have the same skills.
- A job is a bundle of tasks within a firm. As we discuss in Section 1.5, we define a job in our data as an occupation–establishment pair.
- Each job belongs to a *market*, and all jobs in the same market are composed of the same bundle of tasks.
- There are $I$ worker types, indexed by $\iota$, and $\Gamma$ markets, indexed by $\gamma$.
- The key parameter of the model is an $I \times \Gamma$ productivity matrix, $\Psi$, where the $(\iota, \gamma)$ cell, $\psi_{\iota\gamma}$ denotes the number of efficiency units of labor a type $\iota$ worker can supply to a job in market $\gamma$.[7]

Time is discrete, with time periods indexed by $t \in \{1, \ldots, T\}$, and workers make idiosyncratic moves between jobs over time. Neither workers, households, nor firms make dynamic decisions, meaning that the model may be considered one period at a time. We do not consider capital as an input to production. We use the model to (i) microfound our network-based method for assigning workers to worker types and jobs to markets, (ii) identify model parameters, and (iii) quantify the effects of labor market shocks on workers.

### 1.2.2 Household

A representative household consumes output from each sector as inputs to a constant elasticity of substitution (CES) utility function. Utility is given by

$$U = \left( \sum_{s=1}^{S} a_s^{\frac{1}{\eta}} y_s^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}} \tag{1.2.1}$$

where $C$ is a numeraire aggregate consumption good, $y_s$ is the household's consumption of sector $s$'s output, $\eta$ is the elasticity of substitution between sectors' output, and $a_s$ is

[7] We can think of $\psi_{\iota\gamma}$ as $\psi_{\iota\gamma} = f(X_\iota, Y_\gamma)$, where $X_\iota$ is an arbitrarily high dimensional vector of skills for type $\iota$ workers, $Y_\gamma$ is an arbitrarily high dimensional vector of tasks for jobs in market $\gamma$, and $f()$ is a function mapping skills and tasks into productivity. This framework is consistent with Acemoglu and Autor (2011)'s skill and task-based model, and is equivalent to Lindenlaub (2017) and Tan (2018). A key difference is that Lindenlaub and Tan observe $X$ and $Y$ directly and assume a functional form for $f()$, whereas we assume that $X$, $Y$, and $f()$ exist but are latent. We do not identify $X$, $Y$, and $f()$ directly because in our framework $\psi_{\iota\gamma}$ is a sufficient statistic for all of them.

a demand shifter for the sector $s$ good. In our counterfactual analyses we generate labor demand shocks by changing the vector of sector demand shifters $\vec{a}$. It follows that the demand curve for sector $s$'s output is given by

$$y_s^D = \frac{a_s}{\sum_{s'} \left(\frac{p_s}{p_s'}\right)^{\eta} (a_{s'} p_{s'})} Y \tag{1.2.2}$$

where $Y$ is total income.

The household consumes its entire income each period, meaning that $Y = \sum_s p_s y_s^D$. Because all workers belong to the household and the household owns all firms, total income is the sum of all labor income and profits in the economy: $Y = \bar{W} + \Pi$.

### 1.2.3 Firms

There are $S$ sectors indexed by $s$. Each sector $s$ consists of a continuum of firms in a competitive sector-level product market. Each firm, indexed by $f$, has a Cobb-Douglas production function which aggregates tasks from different labor markets, indexed by $\gamma$. The quantity of the sector $s$ good produced by firm $f$, $y_{sf}$, is therefore given by

$$y_{sf} = \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} \tag{1.2.3}$$

where $\ell_{\gamma f}$ is the number of efficiency units of labor firm $f$ employs in jobs in market $\gamma$, and $\beta_{\gamma s}$ is the elasticity of sector $s$ output with respect to labor employed in market $\gamma$ in sector $s$.

The firm chooses labor inputs in order to maximize profits, taking as given the price of output $p_s$, a vector of wages per efficiency unit of labor $w_\gamma$, and a production function, equation (1.2.3). Therefore, the firm solves

$$\pi_f = \max_{\{\ell_{\gamma f}\}_{\gamma=1}^{\Gamma}} \quad p_s \cdot \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} - \sum_{\gamma} w_\gamma \ell_{\gamma f}. \tag{1.2.4}$$

Production exhibits decreasing returns to scale because

$$\sum_{\gamma} \beta_{\gamma s} = \alpha < 1 \quad \forall s$$

where $\alpha$ denotes the labor share.

We define a job, indexed by $j$, as a firm-market pair. Therefore, we can replace the $\gamma f$ indices with $j$ in the equations above: $\ell_{\gamma f} \equiv \ell_j$. We denote the market to which job $j$ belongs

as $\gamma(j)$. It is possible for multiple workers to be employed by the same job at the same time. For example, if "economist" is a market, then "economist at the University of Michigan" would be a job and it would employ approximately 50 workers. Total profits in the economy are the sum of all firms' profits: $\Pi = \sum_{s=1}^{S} \sum_{f \in s} \pi_f$.

### 1.2.4 Workers

Workers, indexed by $i$, are endowed with a *worker type*, indexed by $\iota$, and one indivisible unit of labor. We denote worker $i$'s type as $\iota(i)$. There is an exogenously-determined mass of type $\iota$ workers, $m_\iota$. The worker's type defines their skills. Type $\iota$ workers can supply $\psi_{\iota\gamma}$ efficiency units of labor to jobs in market $\gamma$. $\psi_{\iota\gamma}$ is a reduced form representation of the skill level of a type $\iota$ worker in the various tasks required by a job in market $\gamma$. Units of human capital are perfectly substitutable, meaning that if type 1 workers are twice as productive as type 2 workers in a particular market $\gamma$ (i.e. $\psi_{1\gamma} = 2\psi_{2\gamma}$), firms would be indifferent between hiring one type 1 worker and two type 2 workers at a given wage per efficiency unit of labor, $w_\gamma$. Therefore, the law of one price holds for each market, and a type $\iota$ worker employed in a job in market $\gamma$ is paid $\psi_{\iota\gamma} w_\gamma$. Because workers' time is indivisible, each worker may supply labor to only one market in each period and we do not consider the hours margin.

Workers' only decisions are their market choices. Workers are indifferent between individual jobs in the same market, meaning that individual jobs face perfectly elastic labor supply at the wage for their market, $w_\gamma$.[8] In addition to earnings, each market $\gamma$ has a fixed amenity value to workers, $\xi_\gamma$; $\Xi = \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_\Gamma \end{bmatrix}$. Workers may also choose to be non-employed, denoted by $\gamma = 0$, in which case they receive no wages but receive a non-employment benefit, which is normalized to 0 without loss of generality. Finally, each worker $i$ has an idiosyncratic preference for market $\gamma$ jobs at time $t$, $\varepsilon_{i\gamma t}$. Therefore, worker $i$ chooses a market by solving

$$\gamma_{it} = \underset{\gamma \in \{0,1,\ldots,\Gamma\}}{\arg\max}\ \psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma + \varepsilon_{i\gamma t} \tag{1.2.5}$$

where $\gamma_{it}$ denotes the market worker $i$ chooses to supply labor to at time $t$. We assume that $\varepsilon_{i\gamma t}$ is iid type 1 extreme value with scale parameter $\nu$:

**Assumption 1.2.1** (Distribution of preference shocks)**.** *Idiosyncratic preference shocks $\varepsilon_{i\gamma t}$ are drawn from a type-I extreme value distribution with dispersion parameter $\nu$ and are serially uncorrelated and independent of all other variables in the model.*

[8] If workers do not view all jobs of the same type as identical, then individual jobs would face an upward-sloping labor supply curve, and would thus have some degree of market power. We explore this in Chapter II.

This gives us a functional form for the probability that a type $\iota$ worker chooses a job in market $\gamma$:

$$\mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu] = \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)}. \tag{1.2.6}$$

We aggregate over individual workers to specify labor supply. As noted above, $m_\iota$ denotes the exogenously-determined mass of type $\iota$ workers. The *number* of workers employed in market $\gamma$ jobs is

$$NumWorkers_\gamma(\vec{w}_t) = \sum_\iota m_\iota \mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu] = \sum_\iota m_\iota \left( \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)} \right).$$

The expression above does not correspond to the labor supply curve that clears the market. In order to clear the market, the *quantity* of labor supplied to market $\gamma$ jobs must equal demand. To get the *quantity* of labor supplied to market $\gamma$ jobs, rather than the number of workers, we weight the equation above by the number of efficiency units of labor supplied by a type $\iota$ worker to a job in market $\gamma$: $\psi_{\iota\gamma}$:

$$LS_\gamma(\vec{w}_t) = \sum_\iota m_\iota \mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu]\psi_{\iota\gamma} = \sum_\iota m_\iota \left( \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma} \tag{1.2.7}$$

### 1.2.5 Timing

We observe the economy for $T$ periods. In each period a worker may draw a Poisson-distributed exogenous separation shock, denoted $c_{it} = \mathbb{1}_{j(i,t)\neq j(i,t-1)}$ where $j(i,t)$ is the job employing worker $i$ at time $t$ (Assumption 1.2.2). Workers who draw a separation shock receive a new set of idiosyncratic preference shocks $\varepsilon_{i\gamma t}$ and search again following the same optimization problem defined in equation (1.2.5). We assume that the labor market parameters, $\{\Psi, \Xi, \nu\}$, and the demand shifters $\vec{a}$, are fixed across all $T$ time periods (Assumption 1.2.3).

**Assumption 1.2.2** (Exogenous separations)**.** *Job separations for worker $i$, $c_{it}$, arrive at a worker-specific Poisson rate $d_i$, and are serially uncorrelated and independent of all other*

*variables in the model.*

**Assumption 1.2.3** (Constant parameters)**.** *The labor supply parameters,* $\{\Psi, \Xi, \nu\}$*, are constant over the periods in which we estimate the model and perform counterfactuals. The product demand shifters,* $\vec{a}$*, are constant over the periods in which we estimate the model.*

These restrictions make the model a reasonable approximation for relatively short periods of time, but it would be inappropriate for studying long-run changes when labor supply parameters may be changing.

The timing of the model is as follows. In each period $t$:

1. Each employed worker draws an exogenous separation shock with probability $d_i$; workers who do not receive a separation shock remain in their current job
2. Separated workers receive new preference shocks $\varepsilon_{i\gamma t}$
3. Separated workers choose a market $\gamma_{it}$ according to $\mathbb{P}_\iota[\gamma_{it}|\vec{w}]$
4. Separated workers randomly match with a job within their chosen market $\gamma$

Assumptions 1.2.2 and 1.2.3 allow workers to move between jobs over time, generating the network of worker–job matches that is key to identifying worker types and markets. They also imply that worker movement between jobs is idiosyncratic, meaning that each of a worker's jobs represent i.i.d. draws from the same match probability distribution. We discuss this further in Section 1.3.3.

### 1.2.6 Definition of equilibrium

The model solution consists of vectors of goods prices $\vec{p} := \{p_s\}_{s=1}^S$ and wages per efficiency unit of labor $\vec{w} := \{w_\gamma\}_{\gamma=1}^\Gamma$ that satisfy all equilibrium conditions in each period. Since our model can be solved one period at a time with no cross-time dependence and the fundamentals of the economy are assumed to be constant over our estimation window, the equilibrium conditions below are the same in every period. We solve the model numerically. Our equilibrium has the following components:

1. The labor demand functions $\ell_{\gamma f}$ solve the firms' problem (1.2.4)
2. Labor supply is consistent with workers' expected utility maximization (1.2.6)
3. Goods markets clear. Specifically, demand from the representative household $y_s^D$ equals supply created by evaluating the production function at the optimal level of labor inputs and aggregating over all firms in the sector: $y_s = \sum_{f \in s} \prod_\gamma \ell_{\gamma f}^{\beta_{\gamma s}}$ (1.2.3).

4. The labor market clears for each market $\gamma$: $LS_\gamma = LD_\gamma := \sum_s \sum_{f \in s} \ell_{\gamma f}$

5. Aggregate consumption is equal to income: $Y = \sum_s p_s y_s^D = \bar{W} + \Pi$.

### 1.2.7 Discussion

The matrix

$$\Psi = \begin{array}{c} \\ \iota = 1 \\ \iota = 2 \\ \vdots \\ \iota = I \end{array} \begin{array}{c} \begin{array}{cccc} \gamma = 1 & \gamma = 2 & \cdots & \gamma = \Gamma \end{array} \\ \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1\Gamma} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2\Gamma} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{I1} & \psi_{I2} & \cdots & \psi_{I\Gamma} \end{pmatrix} \end{array} \tag{1.2.8}$$

captures productivity heterogeneity resulting from worker skill–job task match complementarities and is the key parameter of our model. As noted above, the typical element of $\Psi$, $\psi_{\iota\gamma}$, captures the effective units of labor a type $\iota$ worker can supply to a job in market $\gamma$. Therefore, $\Psi$ governs both absolute and comparative advantage. Each row of $\Psi$, $\psi_\iota = \begin{bmatrix} \psi_{\iota 1} & \psi_{\iota 2} & \cdots, & \psi_{\iota\Gamma} \end{bmatrix}$, represents a productivity vector for type $\iota$ workers and is a reduced form representation of their skills.

$\Psi$ embeds a flexible notion of skills. It allows us to say that a particular type of worker is highly skilled *in market* $\gamma$, rather than that a type of worker is highly skilled more generally. For example, it allows for a carpenter to be highly skilled at woodworking and an economist to be highly skilled at causal inference without requiring us to classify either type of worker as high-skill or low-skill in general.

$\Psi$ nests three common assumptions about the nature of worker skills. In the standard representative worker framework, worker types do not differ in terms of their skills, but some markets may be more productive than others. This can be represented as $\psi_{\iota\gamma} = \psi_{\iota'\gamma} = \psi_\gamma$ for all $\iota \neq \iota'$. If worker types are differentiated in their skill level but there are no complementarities between worker skills and job tasks, then workers' skills can be represented by a unidimensional index (worker fixed effects). This can be represented as $\psi_{\iota\gamma} = \psi_{\iota\gamma'} = \psi_\iota$ for all $\gamma \neq \gamma'$. If workers' skills are perfectly specific — each worker type can perform exactly one type of job and skills cannot be transferred to other types of jobs — then $\Psi$ is a square diagonal matrix.

**Representative worker**

$$\Psi = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_\Gamma \\ \psi_1 & \psi_2 & \cdots & \psi_\Gamma \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1 & \psi_2 & \cdots & \psi_\Gamma \end{bmatrix}$$

**Worker fixed effect**

$$\Psi = \begin{bmatrix} \psi_1 & \psi_1 & \cdots & \psi_1 \\ \psi_2 & \psi_2 & \cdots & \psi_2 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_I & \psi_I & \cdots & \psi_I \end{bmatrix}$$

**Specific skills**

$$\Psi = \begin{bmatrix} \psi_{11} & 0 & \cdots & 0 \\ 0 & \psi_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_{I\Gamma} \end{bmatrix}$$

## 1.3 Classifying workers and jobs

In this section we derive our procedure for assigning workers to worker types, $\iota$, and jobs to markets, $\gamma$, from the model described in the previous section. The data we use to classify workers and jobs is the set of all worker–job matches, which is the realization of a random matrix $\mathbf{A}$, known as an *adjacency matrix* in network theory parlance. $\mathbf{A}$ has typical element $A_{ij}$, which represents the number of matches between worker $i$ and job $j$. $A_{ij}$ follows a probability distribution derived from our model that depends upon worker $i$'s worker type, $\iota(i)$, and job $j$'s market, $\gamma(j)$. We use the distribution of $\mathbf{A}$ to define a maximum likelihood estimator that assigns workers to worker types and jobs to markets. The estimator formalizes the intuition that two workers belong to the same worker type, $\iota$, if they have the same vectors of match probabilities over markets, and two jobs belong to the same market, $\gamma$, if they have the same vectors of match probabilities over worker types.

### 1.3.1 Assigning workers to worker types and jobs to markets

As stated in equation (1.2.6), when any worker $i$ belonging to type $\iota$ searches for a job, the probability that they choose a job in market $\gamma$ is

$$\mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu] = \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)}$$

This quantity corresponds to a discrete choice at a specific time, $t$. Our assumption that the labor supply parameters ($\Psi$, $\Xi$, and $\nu$) and demand shifters ($\vec{a}$) are unchanging during our estimation period, combined with the fact that $\vec{w}_t$ is determined in equilibrium by the labor supply parameters and demand shifters, means that this choice probability does not depend on the time period. Therefore, we drop the time subscript $t$ in what follows. All workers make this choice in period 1, and workers subsequently make another choice following this distribution any time they experience an exogenous separation.

The quantity in equation (1.2.6), $\mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu]$, refers to the probability of an

individual worker $i$ matching with *any* job in market $\gamma$, not a particular job $j$. To obtain the probability that worker $i$ matches with a *specific* job $j$ in market $\gamma$, we multiply the choice probability in equation (1.2.6) by the probability that worker $i$ matches with job $j$, conditional on choosing a job in market $\gamma$. Because we have assumed that all jobs in the same type are identical from the perspective of workers, this probability is equal to job $j$'s share of market $\gamma$ employment. Let $d_j$ denote the number of workers employed by job $j$ during our estimation period.[9] Then job $j$'s share of all market $\gamma$ employment can be written

$$\mathbb{P}[j|\gamma] = d_j / \sum_{j' \in \gamma} d^J_{j'}. \tag{1.3.1}$$

Therefore, when worker $i$ of type $\iota$ searches, the probability that the search results in worker $i$ matched with job $j$ is the product of the probabilities in equation (1.2.6) and equation (1.3.1):

$$\mathbb{P}_{ij} = \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)}}^{\mathbb{P}_i[\gamma|\Psi,\vec{w},\Xi,\nu]} \times \overbrace{\underbrace{\frac{1}{\sum_{j' \in \gamma} d^J_{j'}}}_{\substack{1/\text{ type } \gamma \\ \text{employment}}} \times \underbrace{d_j}_{\substack{\text{Job } j \\ \text{employment}}}}^{\mathbb{P}[j|\gamma]}. \tag{1.3.2}$$

The first term represents the probability that worker $i$ chooses market $\gamma$, while the second represents the probability that worker $i$ chooses job $j$ conditional on choosing market $\gamma$. We can rewrite this expression as the product of a term that depends only on the worker's type and job's market, which we denote $\mathcal{P}_{\iota\gamma}$, and a job-specific term $d_j$:

$$\begin{aligned}\mathbb{P}_{ij} &= \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)} \times \underbrace{\frac{1}{\sum_{j' \in \gamma} d^J_{j'}}}_{\substack{1/\text{ type } \gamma \\ \text{employment}}}}^{:=\mathcal{P}_{\iota\gamma}} \times \underbrace{d_j}_{\substack{\text{Job } j \\ \text{employment}}} \\ &= \mathcal{P}_{\iota\gamma} d_j.\end{aligned} \tag{1.3.3}$$

$\mathbb{P}_{ij} = \mathcal{P}_{\iota\gamma} d_j$ denotes the probability that an individual search ends with worker $i$ matched with job $j$, but $A_{ij}$ is the number of times worker $i$ matches with job $j$ across *all* of $i$'s

[9] In network theory parlance, $d_j$ is the *degree* of job $j$.

searches. Since the number of times worker $i$ searches depends on the number of separation shocks they draw from a $Poisson(d_i)$ distribution, we can show that $A_{ij}$ also follows a Poisson distribution:

$$A_{ij} \sim Poisson\left(d_i d_j \mathcal{P}_{\iota\gamma}\right). \tag{1.3.4}$$

For a complete proof, see appendix A.7.

This gives us a functional form for the process generating our observed network, encoded in $A_{ij}$:

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right) = \prod_{i,j} \frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!} \exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{1.3.5}$$

where $\vec{\iota} = \{\iota(i)\}_{i=1}^N$ is the vector assigning each worker to a worker type, $\vec{\gamma} = \{\gamma(j)\}_{j=1}^J$ is the vector assigning each job to a market, $\vec{d_i} = \{d_i\}_{i=1}^N$, $\vec{d_j} = \{d_j\}_{j=1}^J$, and $\mathcal{P}$ is the matrix with typical element $\mathcal{P}_{\iota\gamma}$. Using this, we estimate the worker type and market assignments for all workers and jobs, $\vec{\iota}$ and $\vec{\gamma}$ respectively, using maximum likelihood.

$$\vec{\iota},\vec{\gamma} = \underset{\substack{\{\vec{\iota}=\iota(i)\}_{i=1}^N,\\ \{\vec{\gamma}=\gamma(j)\}_{j=1}^J}}{\arg\max} \prod_{i,j} \frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!} \exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{1.3.6}$$

This problem actually has five sets of parameters: worker and job match frequencies $\vec{d_i}$ and $\vec{d_j}$, the type-specific match probabilities $\mathcal{P}_{\iota\gamma}$, and the worker and market assignments $\vec{\iota}$ and $\vec{\gamma}$. The worker and job match frequencies, $\vec{d_i}$ and $\vec{d_j}$, are directly observable in the data so we use their actual values. Conditional on group assignments, the number of matches between each worker type–market pair is observable, and we use these to compute observed match probabilities, which we use as our estimate of the true probabilities, $\mathcal{P}_{\iota\gamma}$. The worker and market assignments, $\vec{\iota}$ and $\vec{\gamma}$, are the parameters we choose in order to maximize the likelihood.

Equation (3.2.4) assumes that we know the number of worker types and markets *a priori*, however this is rarely the case in real world applications. Therefore we must choose the number of worker types and markets, $I$ and $\Gamma$ respectively. We do so using the principle of minimum description length (MDL), an information theoretic approach that is commonly used in the network theory literature. MDL chooses the number of worker types and markets to minimize the total amount of information necessary to describe the data, where the total includes both the complexity of the model conditional on the parameters *and* the complexity

of the parameter space itself. MDL will penalize a model that fits the data very well but overfits by using a large number of parameters (corresponding to a large number of worker types and markets), and therefore requires a large amount of information to encode it. MDL effectively adds a penalty term in our objective function, such that our algorithm finds a parsimonious model. This method has been found to work well in a number of real world networks (Peixoto, 2013; 2014b; Rosvall and Bergstrom, 2007). See appendix A.4 for greater detail.

Equation (3.2.4) corresponds to the degree-corrected bipartite stochastic block model (BiSBM), a workhorse model in the community detection branch of network theory (see appendix B.2 for details). It defines a combinatorial optimization problem. If we had infinite computing resources, we would test all possible assignments of workers to worker types and jobs to markets and choose the one that maximizes the likelihood in equation (3.2.4), however this is not computationally feasible for large networks like ours. Therefore, we use a Markov chain Monte Carlo (MCMC) approach in which we modify the assignment of each worker to a worker type and each job to a market in a random fashion and accept or reject each modification with a probability given as a function of the change in the likelihood. We repeat the procedure for multiple different starting values to reduce the chances of finding local maxima. We implement the procedure using a Python package called graph-tool. (https://graph-tool.skewed.de/. See Peixoto (2014a) for details.)

### 1.3.2 Visual intuition of the BiSBM

Figure 1.1 panel (a) provides a simplified visual representation of how our model generates a network of worker–job matches. We assume that there are 2 worker types, 3 markets, and matches are drawn from a sample match probability distribution

$$\mathcal{P}_{\iota\gamma} = \begin{array}{c} \begin{array}{ccc} \gamma=1 & \gamma=2 & \gamma=3 \end{array} \\ \begin{pmatrix} 0.3 & 0.5 & 0.2 \\ 0.15 & 0.05 & 0.8 \end{pmatrix} \end{array} \begin{array}{l} \iota=1 \\ \iota=2 \end{array}$$

Dots on the left axis represent individual jobs $j$ and dots on the right axis represent individual workers $i$. Workers belong to one of two worker types ($\iota \in \{1,2\}$) and jobs belong to one of three markets ($\gamma \in \{1,2,3\}$). Lines represent employment contracts between individual workers and jobs. A line connects worker $i$ and job $j$ if $A_{ij} > 0$, while i and j are not connected if $A_{ij} = 0$. Consistent with $\mathcal{P}_{\iota\gamma}$, we see that type $\iota = 1$ workers match with all 3 markets with somewhat similar probabilities, while type $\iota = 2$ workers overwhelmingly match with type $\gamma = 3$ jobs. In our actual data, we observe neither worker types and markets,

nor worker type-market match probabilities. We only observe matches between individual workers and jobs, as represented by $A_{ij}$, and visualized here in panel (b) of Figure 1.1. Therefore, our task, formalized in the maximum likelihood procedure defined in equation (3.2.4), is to take the data represented by panel (b) and label it as we do in panel (a). Intuitively, two workers belong to the same worker type if they have approximately the same vectors of match probabilities over all markets, and two jobs belong to the same market if they have approximately the same vector of match probabilities over all worker types.

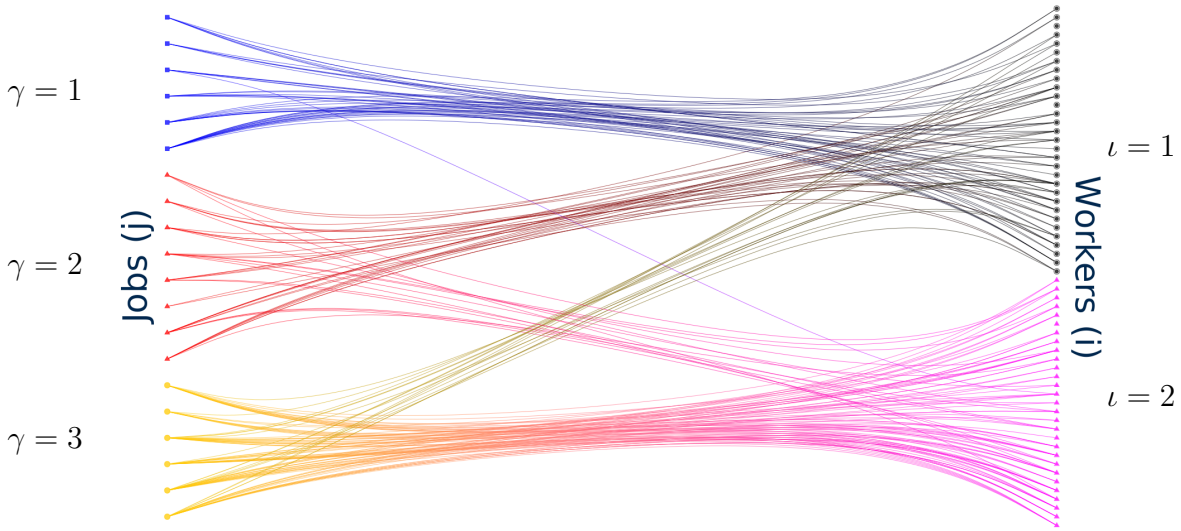
### 1.3.3 Discussion

Our approach rests on the insight that workers with similar propensities to match with particular jobs have similar skills, while jobs with similar propensities to hire particular workers require similar tasks. We formalize this by making three major assumptions. First, our model implicitly assumes that workers match with jobs according to comparative advantage, where comparative advantage is governed by the productivity of the worker's skills when employed in the job's tasks (equation 1.2.6). Second, Assumption 1.2.3 states that the fundamentals of the economy — the labor supply parameters $\Psi$, $\Xi$, and $\nu$, and the demand shifters $\vec{a}$ — are fixed throughout our estimation window. Third, combining the assumptions of i.i.d. T1EV preference shocks (Assumption 1.2.1) and exogenous separations (Assumption 1.2.2), we assume that movement of workers between jobs represents idiosyncratic lateral moves. This allows us to treat a worker's multiple spells of employment as repeated draws from the same distribution, however, as we discuss below, this comes at the cost of ignoring the possibility that workers are climbing the career ladder or that worker flows represent structural shifts in the economy. These assumptions allow us to write the data generating process of the linked employer-employee data in equation (3.2.3), which in turn implies a maximum likelihood estimation strategy. Now, we address the ramifications of these assumptions in turn.

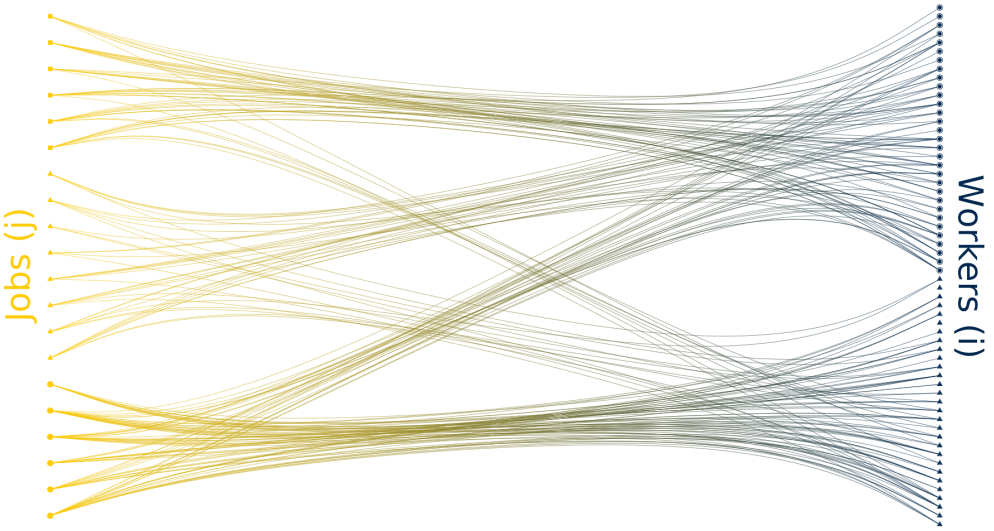
The first major assumption is that workers and jobs match according to a Roy model in which match probabilities are driven by skill-task match productivity. Since workers and jobs are clustered according to match probabilities, to the extent that match probabilities are determined by factors other than skills and tasks, we are clustering on the basis of these other factors. For example, if two groups of workers have very similar skills but rarely end up in the same jobs because they have different credentials, they would be assigned to different worker types, reflecting heterogeneity in credentials rather than skills. Similarly, we may identify groups of workers with similar skills but different preferences. For example, liberal and conservative political consultants may have very similar skills, but consider entirely disjoint sets of jobs due to their preferences. If this is true, our model would assign them

Figure 1.1: Network representation of the labor market

(a)




(b)




Dots represent individual workers/jobs; lines represent employment contracts. Network drawn according to

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right)$$

where

$$P_\iota[\gamma_{it}|\vec{w}]=\begin{array}{c}\begin{matrix}\gamma=1 & \gamma=2 & \gamma=3\end{matrix}\\ \begin{pmatrix}0.3 & 0.5 & 0.2\\ 0.15 & 0.05 & 0.8\end{pmatrix}\end{array}\begin{matrix}\\ \iota=1\\ \iota=2\end{matrix}$$

to different worker types. If there is discrimination, for example on the basis of race or gender, this would be reflected in our productivity measure: our model would assume that certain workers are not being hired because they have low productivity, when in reality they are being discriminated against. Finally, while we restrict to a single metropolitan area to minimize the role of geography, our "skills" and "tasks" may also reflect geographic location and associated commuting costs. Therefore, what we call "skills" should be interpreted more generally as worker characteristics valued by jobs in the labor market, and similarly for "tasks." This is an appealing feature of our method because our agnostic approach to defining labor market relevant worker characteristics allows us to identify clusters of workers who are viewed by the market as approximately perfect substitutes, and these clusters are the relevant units of analysis when considering the effects of shocks on workers. Our method would, however, be inappropriate for studying changes in how worker characteristics are viewed by the market, for example changes in occupational licensing laws or discrimination. A similar logic applies to jobs and tasks.

The second assumption is that the fundamentals of the economy — the assignments of individual workers and jobs to worker types and markets, the labor supply parameters $\Psi$, $\Xi$, and $\nu$, and the demand shifters $\vec{a}$ — are fixed throughout our estimation window. This assumption allows us to identify worker types and markets from the network of worker–job matches. It implies that the network is drawn i.i.d. from an unchanging probability matrix $\mathcal{P}$, meaning that if two workers have the same vector of match probabilities it must be because they have the same vector of skills, and similarly for jobs. The static fundamentals assumption implies that we must estimate the model during a period of time in which the labor market experiences no large shocks.[10] [11]

Finally, we assume exogenous separation shocks in order to rationalize the fact that while worker–job matches are somewhat persistent, we still observe job-to-job transitions even when the fundamentals of the economy are unchanging. We could have alternatively rationalized persistent matches by allowing for endogenous separations alongside persistent idiosyncratic preferences $\varepsilon_{it}$, however exogenous separations are more tractable.[12] An implication of the exogenous separations assumption is that a worker's match probabilities

[10] While we need the demand shifters $\vec{a}$ to be fixed during the estimation window, we may still use our model to estimate the effect of demand shocks if we are able to estimate the parameters during a static pre-shock period and then the shock changes the demand shifters, but not the parameters of the economy, including the worker types, markets, and labor supply parameters.

[11] Endogenously determined wages also drive observed matching patterns, but this is not a problem for our identification strategy. As long as the fundamentals of the economy are fixed, workers of the same type will still display similar matching probabilities and will be clustered together according to our method. In other words, even though the wage distribution shapes the matching patterns in the labor market, similar workers will still behave similarly if fundamentals are fixed.

[12] See Grigsby (2019, Appendix D) for details on this alternative approach.

are independent of their job history, conditional on their type.[13]

## 1.4 Estimating labor supply parameters

This section describes the procedure we use to estimate the labor supply parameters of the model, conditional on the assignments of workers to worker types, $\iota(i)$, and jobs to markets, $\gamma(j)$, described in the previous section.

### 1.4.1 Estimating $\Psi$ from observed matches

Identification and estimation of the labor supply parameters builds upon Bonhomme et al. (2019) and Grigsby (2019), with the key difference being that we assign both workers to worker types and jobs to markets prior to estimating labor supply parameters and do so in a way that more fully exploits the information revealed by worker–job matches, allowing us to identify a significantly greater degree of worker and job heterogeneity.[14]

We estimate parameters using a maximum likelihood approach. We assume that individual workers' earnings in period $t$ are observed with multiplicative measurement error $e_{it}$, which has a worker type–market-specific parametric distribution $f_e(e_{it}|\iota(i), \gamma_{it}, \theta_e)$ with

[13]This rules out job ladders in which the identity of a worker's next job depends on the identity of their current job. We view this as a reasonable approximation for two reasons. First, our model is intended to analyze relatively short periods of time, over which workers skills are fixed and promotions up the career ladder are less frequent. Second, our aim is to identify groups of workers and jobs which are similar in the sense of being substitutable for each other. If one job lies directly above another on the career ladder, meaning that the higher job routinely hires workers from the lower job, then these jobs hire workers with similar skills, and therefore likely require similar tasks. If there was a large increase in employment at jobs on the higher level of the ladder, many of these workers would presumably be hired from jobs at the lower level of the ladder, implying that these workers can reasonably be assigned to the same type. This is effectively a question of whether or not to merge two similar worker types, and we answer it using MDL. However, it would be possible to extend our model to allow for job ladders by modeling the temporal relationship between a worker's multiple job matches.

[14]More precisely, Bonhomme et al. (2019) model workers matching with firms and therefore use k-means clustering to cluster firms on the basis of the firms' earnings distributions, while Grigsby (2019) models workers matching with clusters of occupations identified by combining occupational education requirements with k-means clustering on the basis of occupations' O*NET skills scores. Additionally, neither Bonhomme et al. (2019) nor Grigsby (2019) actually assign workers to types. Instead, they employ random effects estimators, in which they identify the distribution of types, rather than assigning any individual worker to a type. As a result, both papers require that flows of worker types between firm/occupation groups form a strongly connected graph (they use the term "connecting cycle"). This is a strong data requirement and requires them to define worker and firm/occupation groups at a relatively aggregated level, ignoring considerable heterogeneity. By using the network structure of the data to assign workers and jobs to types in a previous step before estimating labor supply parameters, we are able to identify an order of magnitude more worker types and markets, and therefore to allow for much greater heterogeneity.

unit mean, summarized by parameter vector $\theta_e$. Observed earnings $\omega_{it}$ are therefore

$$\omega_{it} = \psi_{\iota(i)\gamma_{it}} w_{\gamma_{it}} e_{it}. \tag{1.4.1}$$

Finally, we assume that the earnings measurement errors are serially independent:

**Assumption 1.4.1** (Serial independence of earnings measurement error)**.** *The realization of period t's measurement error for worker i, $e_{it}$ is independent of the history of errors $\{e_{it'}\}_{t'=1}^{t-1}$, market choices $\{\gamma_{it'}\}_{t'=1}^{t-1}$, and separations $\{c_{it'}\}_{t'=1}^{t-1}$, conditional on the worker's type, $\iota_i$, and current market choice $\gamma_{it}$.*

Our model is identified by combining assumption 1.4.1 with assumptions 1.2.1 and 1.2.2, which stated that the market preference parameters $\varepsilon_{i\gamma t}$ and exogenous separation shocks $c_{it}$ are each serially uncorrelated and independent of all other variables in the model.

Conditional on clustering workers and jobs into types, our data consist of three elements per worker per period: the worker's market choice, $\gamma_{it}$, the worker's earnings, $\omega_{it}$, and the indicator for whether or not the worker changed jobs, $c_{it}$. Observed data are denoted by $\mathbb{X} := \{\gamma_{it}, \omega_{it}, c_{it} | t = 1, \ldots, T; i = 1, \ldots, N\}$. The parameters are denoted by $\Theta := \{\psi_{\iota\gamma} w_\gamma, \xi_\gamma, \nu, \theta_e | \iota = 1, \ldots, I; \gamma = 1, \ldots, \Gamma\}$. Recall that $\mathbb{P}[\gamma_{it}|\Theta]$ is the probability of worker $i$ choosing a job in market $\gamma$ and comes from the Roy model (equation 1.2.6). Meanwhile, let $f_\omega(\omega|\iota(i), \gamma_{it}, \Theta)$ denote the density of observed earnings in period $t$. We construct our likelihood as follows.

In periods in which workers experience a separation, three pieces of data are generated: a separation indicator $c_{it}$, the worker's new market choice $\gamma_{it}$, and the worker's earnings $\omega_{it}$. We assume that all workers separate and rematch in the first period for which we have data: $c_{i1} = 1$ for all $i$. In periods in which the worker does not separate from their job, we observe only $c_{it}$ and $\omega_{it}$.[15] Assumptions 1.2.2 and 1.4.2 tell us that realizations of $\omega_{it}$ and $c_{it}$ are independent, and $\gamma_{it}$ is independent of $\omega_{it}$ conditional on $c_{it}$. Therefore, we write the likelihood of observing $\{\gamma_{it}, \omega_{it}, c_{it}\}$ for an individual worker in period $t$ as

$$l(\gamma_{it}, \omega_{it}, c_{it}|\mathbb{X}) = \underbrace{[f_\omega(\omega_{it}|\Theta)\mathbb{P}(\gamma_{it}|\Theta)]}_{\text{Separation}}^{c_{it}} \underbrace{[f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{No separation}}^{1-c_{it}}$$

[15] By only including the worker's market choice in the likelihood in periods in which a separation has occurred, but assuming that all workers separated in period $t = 1$, we are ensuring that each match enters the likelihood exactly once. This gives all matches equal weight in the likelihood, regardless of match duration. Alternatively, we could have omitted exogenous separations from the model and assumed that workers make a new choice every period. Under this assumption, persistent matches would indicate that the worker has made the same choice repeatedly and we would put greater weight on persistent matches in estimation.

Our assumptions that $\{\gamma_{it}, \omega_{it}, c_{it}\}$ are serially uncorrelated and independent across workers, conditional on the parameters of the data, allow us to write the full likelihood of the data as the product of the individual worker-time likelihoods:

$$\begin{aligned}\mathcal{L}(\Theta|\mathbb{X}) &= \prod_{i=1}^{N}\prod_{t=1}^{T} l(\gamma_{it}, \omega_{it}, c_{it}|\mathbb{X}) \\ &= \prod_{i=1}^{N}\prod_{t=1}^{T} \underbrace{[\mathbb{P}(\gamma_{it}|\Theta) f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{Separation}}^{c_{it}} \underbrace{[f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{No separation}}^{1-c_{it}} \end{aligned} \tag{1.4.2}$$

Finally, the log-likelihood is

$$\ell(\Theta|\mathbb{X}) = \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it} \log \mathbb{P}(\gamma_{it}|\Theta) + \sum_{i=1}^{N}\sum_{t=1}^{T} \log f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta) \tag{1.4.3}$$

In order to maximize this likelihood function, we impose a distributional assumption and a normalization:

**Assumption 1.4.2** (Distribution of measurement error in wages)**.** *$e_{it}$ has a log-normal distribution:* $\ln e_{it} \sim \mathcal{N}(0, \sigma_{\iota\gamma})$.

**Assumption 1.4.3** ($\Psi$ normalization)**.** *The mean productivity level in each market $\gamma$ is normalized to a constant, $k$:*

$$\sum_{\iota} m_\iota \psi_{\iota\gamma} = k \quad \forall\gamma$$

*where $m_\iota$ is the mass of type $\iota$ workers.*

Assumption 1.4.2 assumes that wages follow a log-normal distribution which is worker type-market specific, following Bonhomme et al. (2019) and Grigsby (2019). Assumption 1.4.3 normalizes the $\psi_{\iota\gamma}$ to have a mean equal to some constant $k$ within market.

Identification of $\Psi$ comes from two sources: earnings for all employed workers, and market choices for all workers in period $t = 1$ and workers who receive exogenous separation shocks in periods $t > 1$. Intuitively, $(\iota, \gamma)$ matches that pay more and occur more frequently are revealed to be more productive. The relative weight of earnings and market choices is determined by the inverse of the variances of measurement error in wages and idiosyncratic shocks — if the earnings measurement error $\sigma_{\iota\gamma}$ for a worker type–market pair has a relatively high variance, then estimation puts more weight on choices; if the idiosyncratic preference shocks have a relatively high variance (large $\nu$), estimation puts more weight on earnings. The normalization that the mean skill level in each market equals $k$ (Assumption 1.4.3)

converts the distribution of relative skills into a distribution of skill levels. We choose $k$ to maximize the model's ability to match the observed employment rate.[16]

The parameter governing the variance of non-pecuniary benefits, $\nu$, is identified by workers' choices of markets, $\gamma$. Workers will choose a market that offers their worker type low expected utility (low $\psi_{\iota\gamma} w_\gamma + \xi_\gamma$) when they receive a large preference shock draw for that market. Therefore, if workers frequently choose low expected utility markets, it must be because they frequently draw large preference shocks, indicating that the preference shock distribution has a large dispersion parameter, $\nu$. The market amenities parameter $\xi_\gamma$ is a market fixed effect and is identified by the component of the frequency with which workers choose market $\gamma$ that is common across all worker types $\iota$. The relative value of $\xi_\gamma$ to $\xi_{\gamma'}$ allows the model to match the fact that some high-earning markets, such as doctors, account for a small share of total employment. This is because $\xi_\gamma$ reflects not just the immediate utility benefits of working in a job in market $\gamma$, but also reflects broader compensating differentials. In this way, $\xi_{doctor}$ may be low, not because doctor jobs are unpleasant, but because the annualized cost of becoming a doctor — including medical school — and maintaining the requisite skills is high. We provide greater detail on identification in appendix A.5.

### 1.4.2 Additional parameters to be estimated or calibrated

We also have the following parameters to estimate or calibrate:

- $\beta_{\gamma s}$ (output elasticity of labor in market $\gamma$) — We calibrate these parameters as the share of the sector $S$ wage bill paid to workers employed in market $\gamma$ jobs.
- $\eta$ (CES consumption substitution elasticity)— We calibrate this parameter to 2.[17]
- $a_s$ (demand shifters) — We calibrate demand shifters to match actual sector output shares, given sector-level prices, for the state of Rio de Janeiro as measured by the Brazilian Institute of Geography and Statistics (IBGE).

[16]This normalization is mostly without loss of generality. If one were to double the number of efficiency units of labor each worker supplied to a market, the equilibrium price of labor would halve. However, increasing the number of efficiency units of labor in the economy will impact the fraction of the labor force in employment versus non-employment. This is why we choose $k$ to maximize the model's ability to match the observed employment rate.

[17]Broda and Weinstein (2006) estimate this parameter to be 4, however their estimate comes from significantly more disaggregated product categories, so we choose a smaller value. This parameter affects our structural results in Section 1.7, but does not affect the reduced form estimates in Section 1.8.

### 1.4.3 Discussion

The worker type–market productivity matrix $\Psi$ captures high-dimensional two-sided (worker and job) heterogenity. It is high-dimensional in the sense that workers' skills and jobs' tasks may have arbitrarily high dimensions, and $\Psi$ serves as a sufficient statistic for the quality of the match between a worker's skills and a job's tasks.

This paper contributes to a growing literature which models worker–job (or worker–firm) matching with two-sided (worker and job) heterogeneity. In order to summarize high-dimensional skill and task heterogeneity, much of this literature estimates a matrix analogous to our $\Psi$. In order to do so, researchers identify clusters of similar workers and clusters of similar jobs using observable worker and job characteristics. For example, Lindenlaub (2017) imputes worker skill groups using information on workers' training and educational degrees, and defines occupation groups using skill requirement information from O*NET. Similarly, Tan (2018) identifies bins of worker skills and job tasks using the ASVAB and O*NET, respectively.[18] These approaches represent imperfect solutions for at least two reasons. First, available measures may measure the skills and tasks valued by the labor market with considerable error. As Frank et al. (2019) note, "according to O*NET, the skill 'installation' is equally important to both computer programmers and to plumbers, but, undoubtedly, workers in these occupations are performing very dissimilar tasks." Second, in many administrative data sets like the LEHD or US income tax data, variables like education, occupation, and direct skill/task measures are not available.[19] Therefore, researchers must resort to survey data, which may have more detailed worker and job characteristics, but have much smaller sample sizes and therefore are unable to capture the level of detailed heterogeneity that we do.[20] While this paper focuses on labor market shocks, our worker classifications can serve as a foundation for future research on worker–job matching or polarization.

Occupation may seem like a solution to this problem, however it too is an imperfect measure of a worker's skills. Workers frequently change occupations without significantly changing their skill sets. In our data, 73 percent of job changes in our data involve changes in occupations (Table 1.2). Moreover, many different occupations require very similar skills. For example, suppose it is the case that retail sales and fast food occupations require similar

[18] Other papers employing a similar framework include Autor et al. (2003); Acemoglu and Autor (2011); Kantenga (2018)

[19] In concurrent work, we are applying our method for classifying workers in order to impute occupation on the LEHD.

[20] Another recent approach to characterizing labor market heterogeneity uses compilations of job postings or resumes, but this literature still faces the challenge of how to aggregate workers and jobs into groups, and our method may help solve the problem.

skills and workers frequently move back and forth between these jobs. Our method would recognize these mobility patterns and cluster these workers as the same worker type. While this example concerns aggregating similar occupations, our method can also be useful for *disaggregating* heterogeneous workers employed in the same occupation. Sticking with the same example, retail sales workers at a specialized luxury retailer may have different skills and perform different tasks than retail sales workers at a discount store. If our data reveal two different clusters of employment relationships — one centered around fast food and discount retail, and the other centered around luxury retail — then our method would recognize this and yield worker types that improve upon classifications based upon occupations by more precisely identifying groups of workers with similar skills. We provide evidence that we succeed in satisfying this objective in Section 1.6. A similar logic applies to clustering jobs into types.

## 1.5 Data

We use the Brazilian linked employer-employee data set RAIS, which contains detailed data on all employment contracts in the Brazilian formal sector. Each observation in the data set represents a unique employment contract and includes a unique worker ID variable, an establishment ID, an occupation code, and earnings. Our sample includes all workers between the ages of 25 and 55 employed in the formal sector in the Rio de Janeiro metro area at least once between 2009 and 2012, We exclude public sector and military employment because institutional barriers make flows between the Brazilian public and private sectors rare. We also exclude the small number of jobs that do not pay workers on a monthly basis.

We create two different analysis data sets — one for classifying workers and jobs using the network of worker–job matches, and one for estimating labor supply parameters ($\Psi$, $\Xi$, and $\nu$) and estimating the effects of shocks on workers. Our data for classifying workers and jobs starts with the sample described above. We define a job as an occupation–establishment pair and generate a unique "Job ID" for each job by concatenating the establishment ID code and the 4-digit occupation code. For example, a job would be "economist at the University of Michigan" and this job would at any given time employ approximately 50 workers. Although we use occupation to define jobs, we do not use occupation as an input to our algorithm for classifying workers and jobs.[21] This gives us a set of worker–job pairs that

[21]For example, we use occupation to assign lawyers and economists at the University of Michigan into separate jobs, but our algorithm does not know that the jobs "Economist at Michigan" and "Economist at the Federal Reserve" correspond to the same occupation. It would only assign these jobs to the same market if they are revealed to be similar by the network of workerjob matches.

define the bipartite labor market network[22] that we use to cluster workers into worker types and jobs into markets. We restrict to jobs employing at least 5 unique workers during our estimation window, though the 5 workers need not be employed by the job simultaneously. This restriction eliminates jobs that are not sufficiently connected to the rest of the network of worker–job matches to infer their match probabilities and assign them to markets.

Once we have assigned workers to worker types and jobs to markets, we create a balanced panel of workers with one observation per worker per year. Our earnings variable is the real hourly log wage in December, defined as total December earnings divided by hours worked. We deflate earnings using the CPI. We exclude workers who were not employed for the entire month of December because we do not have accurate hours worked information for such workers. If a worker is employed in more than one job in December, we keep the job with greater hours. If the worker worked the same number of hours in both jobs, we pick the job with the greatest earnings. If tied on both, we choose randomly. We also merge on each worker's worker type and each job's job type. Workers who are not matched with a job are defined as matching with the outside option, denoted $\gamma = 0$, which includes non-employment and employment in the informal sector.

The RAIS data cover only the formal sector of the Brazilian economy. Therefore, we cannot distinguish between non-employment and informal employment and our outside option, denoted $\gamma = 0$, includes both non-employment and informal sector employment. In 2019, 32.1% of employment in the Rio de Janeiro metropolitan area was in the informal sector.[23] However, transitions between the formal and informal sectors are relatively rare: during our sample period, in a given year, fewer than 2% of formal sector workers moved to the informal sector, and approximately 10% of informal sector workers moved to the formal sector.[24]

We calibrate demand shocks using annual data on real output per sector for the state of Rio de Janeiro from the Brazilian Institute of Geography and Statistics (IBGE). These data are available for 15 sectors, the most disaggregated sector definitions for which annual state-level data are available. The 15 sectors are listed in Table 1.1.

[22] A bipartite network is a network whose nodes can be divided into two disjoint and independent sets $U$ and $V$ such that every edge connects a node in $U$ to a node in $V$. In our case $U$ is the set of workers and $V$ is the set of jobs.

[23] IBGE INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATSTICA. Indicadores de subutilizao da fora de trabalho e de informalidade no mercado de trabalho brasileiro. Rio de Janeiro: IBGE, 2019.

[24] See Engbom et al. (2021, Figure 21).

Table 1.1: IBGE Sectors

| | Sector name |
|---|---|
| 1 | Agriculture, livestock, forestry, fisheries and aquaculture |
| 2 | Extractive industries |
| 3 | Manufacturing industries |
| 4 | Electricity and gas, water, sewage, waste mgmt and decontamination |
| 5 | Construction |
| 6 | Retail, Wholesale and Vehicle Repair |
| 7 | Transport, storage and mail |
| 8 | Accommodation and food |
| 9 | Information and communication |
| 10 | Financial, insurance and related services |
| 11 | Real estate activities |
| 12 | Professional, scientific and technical, admin and complementary svcs |
| 13 | Public admin, defense, educ and health and soc security |
| 14 | Private health and education |
| 15 | Arts, culture, sports and recreation and other svcs |

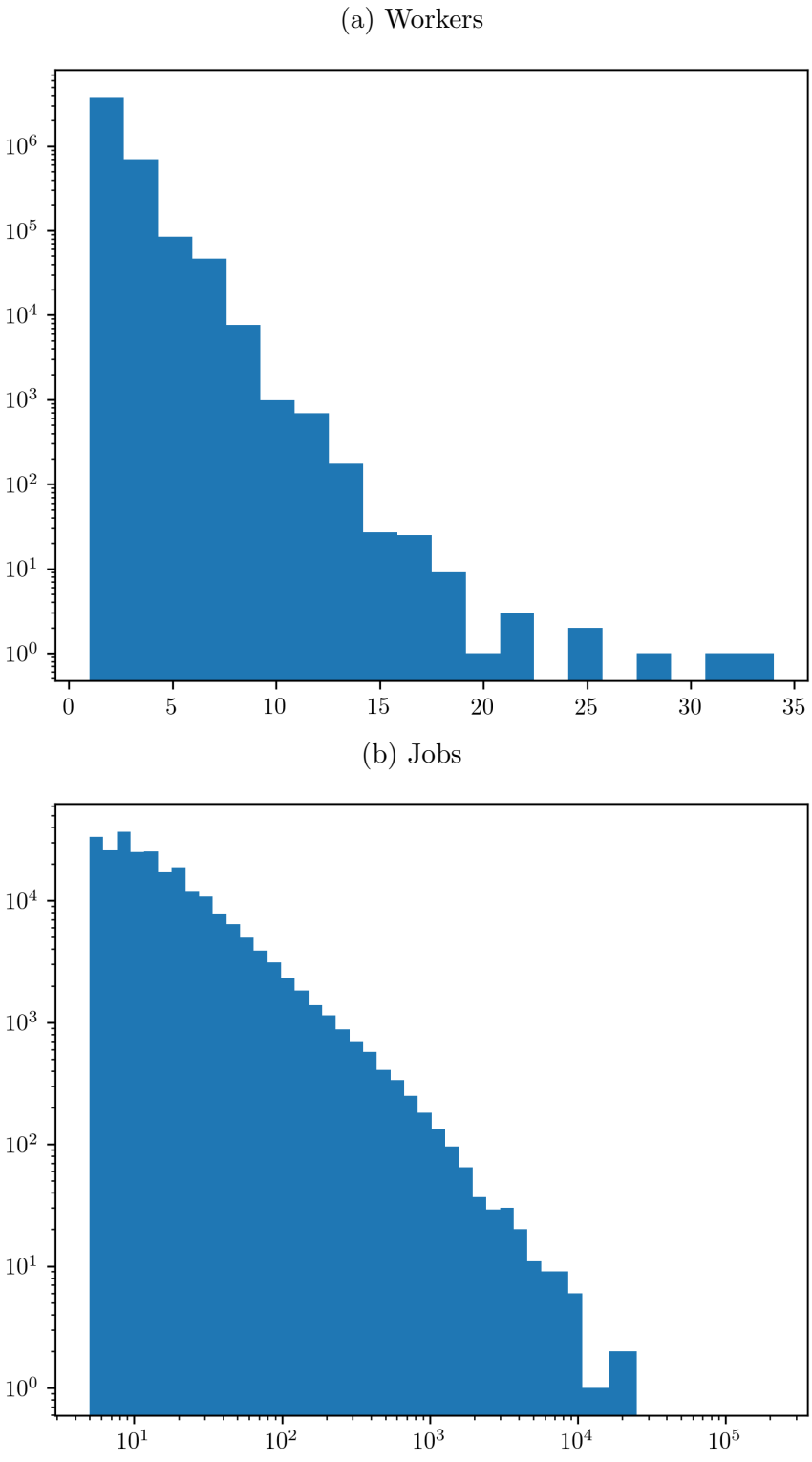
### 1.5.1 Summary statistics

Our data contain 4,578,210 unique workers, 289,836 unique jobs, and 7,940,483 unique worker–job matches. The average worker matches with 1.73 jobs and the average job matches with 27.4 workers. 42% of workers match with more than one job during our sample. Figure 2.1 presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions.

Table 1.2 presents the fraction of job changes that also involve a change in occupation, sector, market, firm, or establishment. The column "All Job Changes" computes the probability that a worker changes occupation, industry, sector, market, firm, or establishment conditional on changing jobs. The column "Firm Change Only" presents the same quantities restricting to the set of job changes that also involve a change in firm. The column "No Firm Change" restricts to job changes that do *not* involve a change in firm. Recall that we define a job as a 4-digit occupation–establishment pair. Table 1.2 shows that 65% of job changes also involve a change in establishment and 54% change firm. This tells us that job changes are not dominated by workers "climbing the job ladder" by changing occupations within a firm.

Table 1.2 also shows that job changes are frequently associated with occupation, industry, and sector changes. 41% of job changes involve a change in 1-digit occupation (most aggregated) and 73% involve a change in 6-digit occupation (most disaggregated). Since

Figure 1.2: Distributions of Number of Matches Per Worker and Job




*Notes:* Figure presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions. Vertical axes presented in log scale. Horizontal axis of bottom panel also presented in log scale. Number of matches per worker and job computed from the network of worker–job matches described in Section 1.5.

occupation, industry, and sector changes are so frequent, it is unlikely that any of these variables precisely measure workers' skills, since workers' skills are unlikely to evolve so quickly. Similarly, the fact that job transitions frequently (59% of the time) involve moving to a job in a different market ($\gamma$) as the old job demonstrates the value of allowing workers to costlessly change the market to which they supply labor, a feature that our model incorporates.

Table 1.2: Occupation/Sector/Market Transition Frequencies

| Variable | All Job Changes | Firm Change Only | No Firm Change |
|---|---|---|---|
| 1-digit Occupation | 0.410 | 0.345 | 0.484 |
| 2-digit Occupation | 0.496 | 0.422 | 0.580 |
| 4-digit Occupation | 0.676 | 0.563 | 0.807 |
| 6-digit Occupation | 0.725 | 0.648 | 0.814 |
| 5-digit Industry | 0.418 | 0.708 | 0.083 |
| Sector (IBGE) | 0.262 | 0.456 | 0.039 |
| Market ($\gamma$) | 0.591 | 0.727 | 0.434 |
| Firm | 0.536 | 1.000 | 0.000 |
| Establishment | 0.645 | 0.996 | 0.240 |

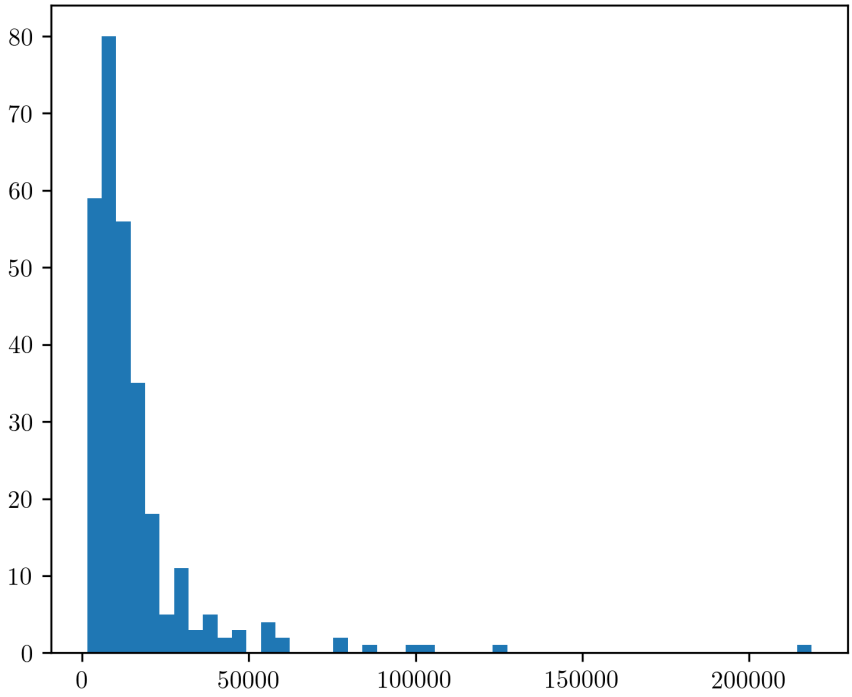
*Notes:* This table presents the fraction of job changes that also involve a change in occupation, sector, market, firm, or establishment. The column "All Job Changes" computes the probability that a worker changes occupation, industry, sector, market, firm, or establishment conditional on changing jobs. The column "Firm Change Only" presents the same quantities restricting to the set of job changes that also involve a change in firm. The column "No Firm Change" restricts to job changes that do *not* involve a change in firm. Since the fraction of job changes that involve a firm change is 0.536, values in the column "All Job Changes" equal 0.536 $\times$ "Firm Change Only" + (1-0.536) $\times$ "No Firm Change." 5-digit sectors refer to narrow industry codes, while there are 15 IBGE sectors, defined in Table 1.1, taken from the Brazilian Institute of Geography and Statistics (IBGE). Values computed using the worker earnings panel described in Section 1.5 using RAIS data from 2009–2012.

## 1.6 Descriptive results

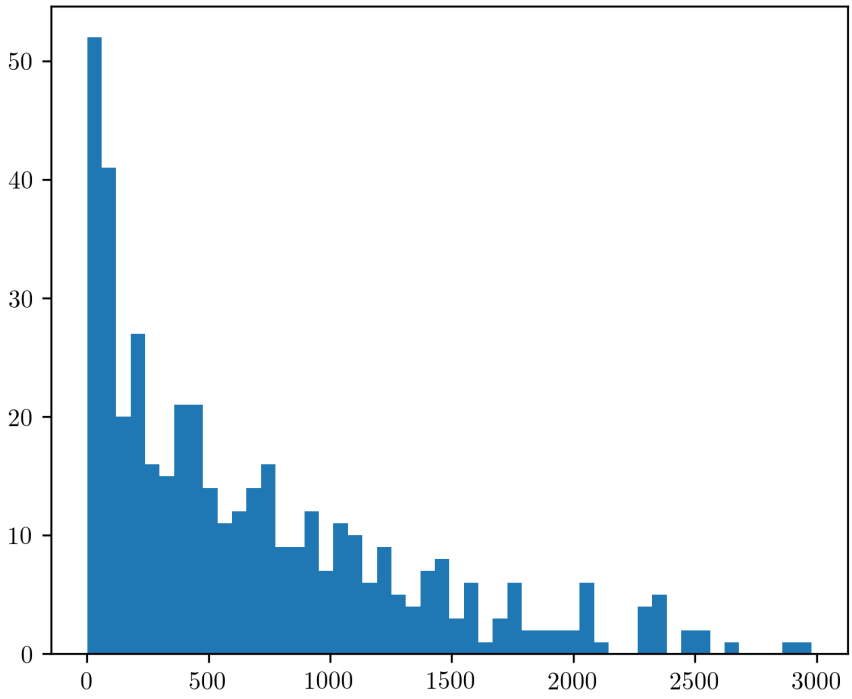
Our network-based classification algorithm identifies 290 worker types ($\iota$) and 427 markets ($\gamma$). Figure 2.2 presents histograms of the number of workers per worker type and jobs per market. The average worker belongs to a worker type with 40,978 workers and the median worker belongs to a worker type with 20,413 workers. The average job belongs to a market with 1,273 jobs and the median job belongs to a market with 1,188 jobs.

Figure 1.3: Worker Type ($\iota$) and Market ($\gamma$) Size Distributions

(a) Number of Workers Per Worker Type ($\iota$)




(b) Number of Jobs Per Market ($\gamma$)




*Notes:* Figure presents histograms of the number of workers per worker type $\iota$ and jobs per market ($\gamma$). The units of analysis are worker types in the upper panel and markets in the lower panel. Computed using assignments of workers to worker types and jobs to markets as described in Section 1.3.

### 1.6.1 Occupation count tables

Our method simultaneously clusters together workers in different occupations who are revealed by the network structure of the labor market to have similar skills, *and* disaggregates workers employed in the same occupation who are revealed to have different skills. As a concrete example, consider the occupation identified by the code 3331-10 in the Brazilian occupation classification system. This occupation is called[25] "course instructor" and is described as

> **Summary description**
> The professionals in this occupational family must be able to create and plan courses, develop programs for companies and clients, define teaching materials, teach classes, evaluate students and suggest structural changes in courses.

Despite this being the most disaggregated level of the occupation classification system (6-digit), there may be considerable heterogeneity within this occupation. This occupation may include, for example, both math tutors and personal fitness trainers — two sets of workers with very different skills. At the same time, it is not obvious what distinguishes a course instructor from a personal trainer (occupation code 2241-20) or an elementary school teacher (occupation code 2312-10). However, if we can identify a cluster of course instructors who at other times in their career work as personal trainers and another cluster who have also worked as elementary school teachers, then we can simultaneously *disaggregate* course instructors with distinct skills, and *aggregate* them by combining them with other workers in different occupations who have similar skills. We pursue these examples in Tables 1.3 and 1.4.

Table 1.3 presents the 10 occupations in which workers belonging to worker type $\iota = 17$ are most frequently employed. To interpret this table, recall that we have assigned each individual worker to a worker type, $\iota$. Each worker may be employed by one or more jobs in our sample, and each job is assigned an occupation code by the Brazilian statistical agency. A worker who has multiple jobs during the sample may have a different occupation associated with each job. This table tabulates how frequently a type $\iota = 17$ worker is employed in each occupation. Most of these occupations are related to physical fitness, education, or both. The most frequently occurring occupation is course instructor. It is not obvious based on the occupation description alone what skills course instructors possess, however because the network structure of the data informs us that these workers have similar skills to personal

[25] Occupation names and descriptions are translated from Portuguese using Google Translate and some translations are imprecise, although manual inspection of a subset by our Portuguese-speaking coauthor confirms that most translations are satisfactory.

trainers, physical education teachers, and sports coaches, it is likely that these workers have skills more closely related to physical education than math.

Now consider Table 1.4. Course instructor is the second most frequently-occurring occupation among type $\iota = 52$ workers, however the other frequently-occurring occupations are teachers of more traditional academic subjects. If we had relied upon occupation codes alone, we would have assumed that all course instructors have the same skills, whereas our clustering approach tells us that there are at least two different types of course instructors: physical education and academic education.

In addition to disaggregating workers in the same occupation with different skills, these tables display our in aggregating workers in different occupations with similar skills. For most of the occupations in these tables, it makes intuitive sense that they should be clustered together. For example, it is not surprising that physical education teachers, sports coaches, and personal trainers would have similar skills. Relying on occupation codes — even the highly-aggregated two-digit occupation codes — would not have grouped these workers together. More generally, we view the fact that our worker types imperfectly align with occupation codes as suggestive evidence of our success in identifying groups of workers with similar skills. Workers with similar skills are likely to be employed in similar occupations, so it would be concerning if our worker types did not overlap with occupations. However, the fact that they only partially overlap with occupations suggests that they capture important dimensions of worker heterogeneity that occupations miss. We develop this argument further in the rest of the paper.

### 1.6.2 Worker type skill correlations

While Section 1.6.1 provided a qualitative example of our method's success in identifying clusters of workers with similar skills, we now provide quantitative evidence of our success in this regard. An ideal worker skills classification scheme will maximize the variance in skills across different worker classifications and minimize the variance of skills within a worker classification. While we do not directly observe individual-level skills and therefore cannot measure within-classification skills variance, we do have a measure of across-classification skills variation. Each element of $\Psi$ represents the productivity of a type $\iota$ worker employed in market $\gamma$. Therefore, $\psi_{\iota\gamma}$ is a summary measure of a type $\iota$ worker's skill at jobs in market $\gamma$, and a full row vector of $\Psi$, $\psi_{\iota\cdot}$, summarizes a type $\iota$ worker's skills in *all* markets. This yields a natural metric for skill similarity across worker types: two worker types, $\iota$ and $\iota'$, have similar skills if their associated productivity vectors $\psi_{\iota\cdot}$ and $\psi_{\iota'\cdot}$ are highly correlated.

If we have done a good job of clustering workers with similar skills into the same type, then the correlations of skills across different worker types will be low. To understand this,

Table 1.3: Top Ten Occupations for Worker Type $\iota = 17$

| Occ-6 | Occupation Name | Share |
|---|---|---|
| 333110 | Course Instructor | .15 |
| 224120 | Personal trainer | .11 |
| 231315 | Physical Education Teacher in Primary School | .08 |
| 224125 | Coach (except for soccer) | .06 |
| 234410 | Physical Education Teacher in Higher Education | .05 |
| 224105 | Fitness monitor | .05 |
| 333115 | Teacher (with High School degree) | .05 |
| 234520 | Education Teacher (with College degree) | .03 |
| 371410 | Recreational Activities Coordinator | .03 |
| 377105 | Professional Athlete (various modalities) | .02 |

*Notes:* Table reports the 6-digit occupations in which workers assigned to worker type $\iota = 17$ are most frequently observed, showing only the 10 most frequent. Values computed using the worker earnings panel described in Section 1.5 using RAIS data from 2009–2012. Occupation classification codes defined according to the Brazilian occupation classification system, *CBO 2002: Classificacao Brasileira de Ocupacoes* and translated from Portuguese to English using Google Translate.

Table 1.4: Top Ten Occupations for Worker Type $\iota = 52$

| Occ-6 | Occupation Name | Share |
|---|---|---|
| 331205 | Elementary School Teacher | .07 |
| 333110 | Course Instructor | .07 |
| 231210 | Elementary School Teacher (1st to 4th grade) | .06 |
| 231205 | Young and Adult Teacher teaching elementary school content | .06 |
| 232115 | High School Teacher | .05 |
| 234616 | English Teacher | .04 |
| 333115 | Teacher of Free Courses | .03 |
| 231305 | Elementary School Science and Math Teacher | .03 |
| 331105 | Kindergarten Teacher | .02 |
| 231310 | Art Teacher in Elementary School | .02 |

*Notes:* Table reports the 6-digit occupations in which workers assigned to worker type $\iota = 52$ are most frequently observed, showing only the 10 most frequent. Values computed using the worker earnings panel described in Section 1.5 using RAIS data from 2009–2012. Occupation classification codes defined according to the Brazilian occupation classification system, *CBO 2002: Classificacao Brasileira de Ocupacoes* and translated from Portuguese to English using Google Translate.

consider an extreme example in which workers were clustered randomly. In this case, all clusters would have exactly the same skills — because the skills of each cluster would just be the average skills of the entire population — and all pairs of productivity vectors would be perfectly correlated. That is, $corr(\psi_{\iota\cdot}, \psi_{\iota'\cdot}) \approx 1$ for all $\iota, \iota'$. Alternatively, we might have two clusters of worker types — for example those intensive in manual skills and those intensive in cognitive skills — such that worker types in the same cluster have highly-correlated skills and those in different clusters have negatively correlated skills. At the other extreme, if skills were perfectly specific (meaning that $\Psi$ was close to a diagonal matrix), skill correlations would be close to zero.

We summarize the correlations between different worker types' productivity vectors in Figure 1.4. We do this in two ways. In the left column we present correlation coefficients between all pairs of the $I = 290$ worker types in a lower triangular $290 \times 290$ matrix (the upper triangular portion is redundant and therefore omitted). Dark red points represent large positive correlations, dark blue points represent large negative correlations, and lighter colors represent smaller correlations. Worker types are sorted by mean earnings, from smallest to largest. In the right column, we present histograms of the correlation coefficients in the left column, along with the standard deviation of the correlation coefficients. The first row presents correlations in which workers are classified by worker type and jobs by market. We provide context for these figures by repeating this exercise using versions of $\hat{\Psi}$ in which workers and jobs are classified using the standard labels in the data: occupation and sector. To do this, we estimate a different version of $\Psi$ using the same maximum likelihood estimation described in Section 1.4, except we classify workers and jobs by occupation and sector, rather than worker type and market. Row 2 of Figure 1.4 shows workers classified by 4-digit occupation and jobs by sector. Row 3 shows workers classified by four-digit occupation and jobs by market ($\gamma$). We choose 4-digit occupations as our primary "status quo" benchmark to compare our method to because occupations are a frequently-used measure of granular worker heterogeneity and because the number of 4-digit occupations in our data (306) is similar to the number of worker types (290), allowing for comparisons at a similar level of granularity.

Figure 1.4 shows that correlations between different worker types' productivity vectors are smaller in magnitude when we use our model's $(\iota, \gamma)$ classifications rather than classifications based on labels available in the data, occupation and sector. This is because the network-based clusters of workers are more successful at segregating workers with distinct skills than are standard occupations. Connecting this to the example in the previous section, if high school and middle school math teachers have similar skills but are classified as distinct worker types, we would observe large correlations (dark red) between their productivity

vectors. By contrast, our worker types disentangle teachers into physical education teachers — including coaches and personal trainers — and teachers in traditional academic subjects. Physical education and academic teachers have less correlated skills than do elementary and middle school teachers. Because we have done a better job of segregating workers with disparate skills, and aggregating workers with similar skills, we observe fewer clusters of highly-correlated worker types.

### 1.6.3 Worker types' labor market concentration

If our model is correct that worker–job matching is largely determined by skill–task match productivity, and we have done a good job of clustering together workers with similar skills and jobs with similar tasks, then each worker type will be concentrated within specific markets. While there will be considerable variation across worker types — worker types with more specific skills will be more concentrated in a small set of markets than those with more general skills — if we compare two job classification schemes, the one that does a better job of identifying workers with similar skills and jobs requiring similar tasks will yield higher worker concentrations in markets.

We compute each worker type's employment concentration across sectors and markets using the Herfindahl-Hirschman index (HHI):

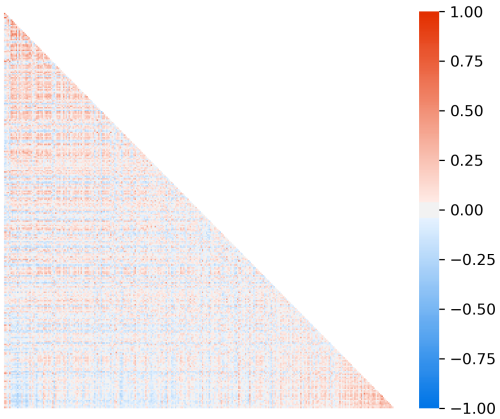
$$HHI_{\iota}^{Sector} = \sum_{s} \pi_{\iota s}^{2} \quad \text{and} \quad HHI_{\iota}^{Market} = \sum_{\gamma} \pi_{\iota \gamma}^{2}$$

where $s$ indexes sectors, $\gamma$ indexes markets, and $\pi_{\iota s}$ and $\pi_{\iota \gamma}$ are the share of type $\iota$ workers employed in sector $s$ and market $\gamma$, respectively. An HHI close to 0 indicates that type $\iota$ employment is spread approximately evenly across sectors/markets, while an HHI close to 1 indicates that type $\iota$ employment is very concentrated in a single sector/market. Suppose we classified jobs randomly. Then worker types would not have a comparative advantage in specific markets and therefore would not be concentrated in specific markets. In this case, the HHI for each worker type would converge to $1/NumJobClassifications$, indicating a uniform distribution of employment across job classifications. At the other extreme, if each worker type had perfectly specific skills and supplied all of its labor to exactly 1 job classification, the HHI would be 1. While we would not expect perfectly specific skills, larger HHIs are evidence that we have done a better job of classifying similar jobs, whereas smaller HHIs imply that we are closer to simply classifying jobs randomly.

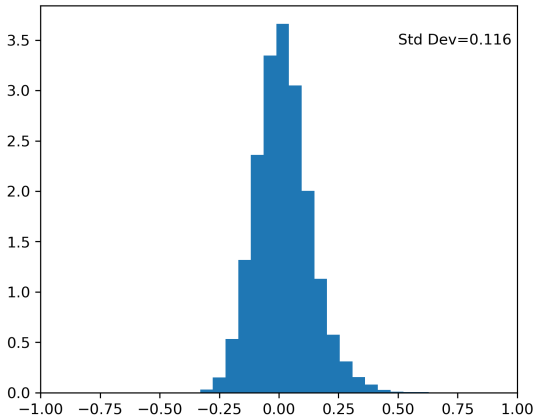
Figure 1.5a presents $HHI_{\iota}^{Sector}$ and $HHI_{\iota}^{Market}$ for each worker type, sorted from least concentrated to most concentrated. Most worker types' labor supply is more concentrated among markets than among sectors, which according to the argument above, indicates that

Figure 1.4: Skill Correlation Across Worker Types and Occupations

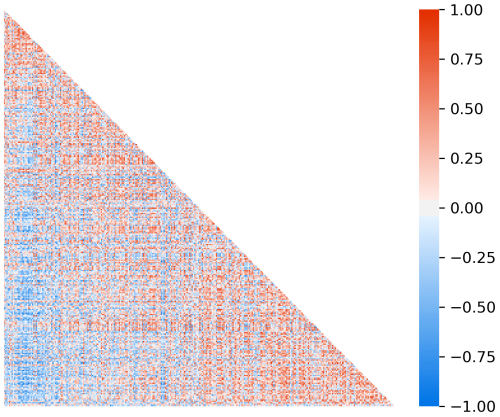
(a) $(\iota, \gamma)$ correlogram



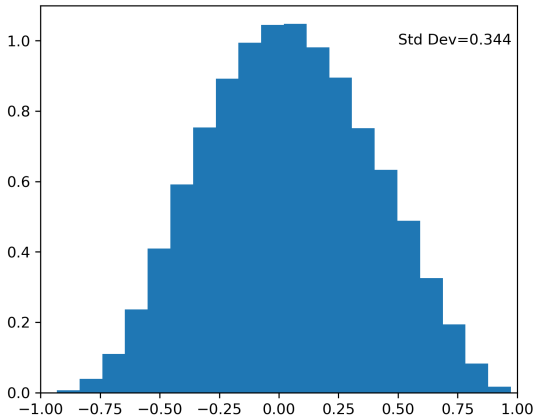

(b) $(\iota, \gamma)$ histogram



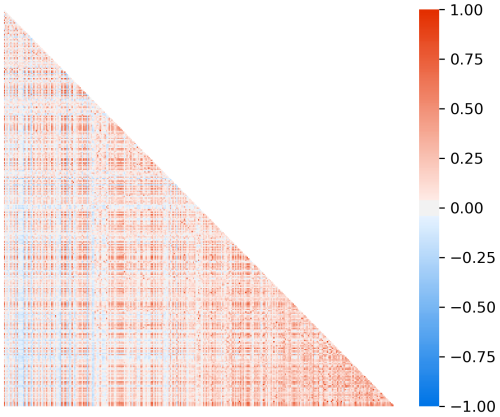

(c) (Occ4, Sector) correlogram



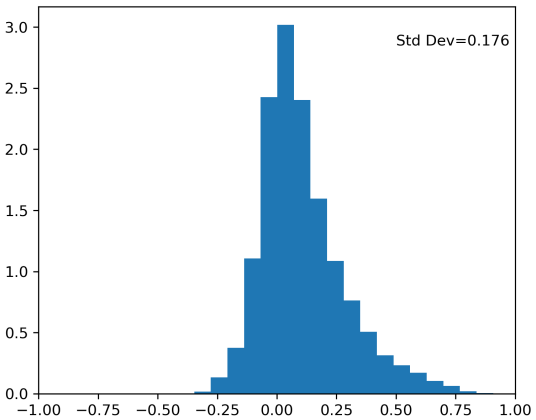

(d) (Occ4, Sector) histogram




(e) (Occ4, $\gamma$) correlogram




(f) (Occ4, $\gamma$) histogram




*Notes:* Figure presents pairwise skills vector correlations (left column) and histograms of these skill correlations (right column) for all pairs of worker types $\iota$ (row 1) and 4-digit occupations (rows 2 and 3). In the left column, dark red squares indicate large positive correlations, while dark blue squares represent large negative correlations. "Skills" defined as row vectors of the matrix $\Psi$, $\psi_{\iota\cdot}$, where $\Psi$ is estimated as described in Section 1.4.1 using the 2009-2012 RAIS worker earnings panel described in Section 1.5. Workers classified by worker types $\iota$ in row 1 and by 4-digit occupation in rows 2 and 3. Jobs classified by market $\gamma$ in rows 1 and 3, and by sector in row 2. Figures in the left column are sorted by worker type mean earnings (smallest to largest).

markets identify groups of jobs that have more homogenous tasks than do sectors. One might be concerned that this isn't a fair comparison because we have 427 markets and only 15 sectors, however having a smaller number of groups mechanically leads to larger HHIs, so this bias runs against the result we find. Nevertheless, in Figure 1.5b we repeat the analysis replacing our 15 sectors with 643 5-digit industries. The qualitative story is the same, but the market HHIs are even larger relative to the industry HHIs than before.

## 1.7 General equilibrium effects of Rio de Janeiro Olympics

We test our model's ability to predict the effects of shocks in the context of the infrastructure investment and other preparations for the 2016 Rio de Janeiro Olympics. The Olympics were announced in late 2009 and construction of new venues and infrastructure were in full effect by 2014. Therefore, we define 2009 as our pre-shock period and 2014 as our "shock" period. We calibrate demand shifters $\vec{a}^{2009}$ and $\vec{a}^{2014}$ to fit sector-level product output in those years, feed these demand shifters through our model and solve for the equilibrium to compute model-implied earnings for each worker type for each year, $\hat{y}_\iota^{2009}$ and $\hat{y}_\iota^{2014}$, and then take the difference $\Delta\hat{y} = \hat{y}_\iota^{2014} - \hat{y}_\iota^{2009}$. We also compute the *actual* mean earnings changes for each worker type, $\Delta y = y_\iota^{2014} - y_\iota^{2009}$. Finally, we regress actual changes in mean earnings on model-predicted changes in mean earnings for each worker type.

$$\Delta y = \beta_0 + \beta_1 \Delta\hat{y} + \varepsilon \tag{1.7.1}$$

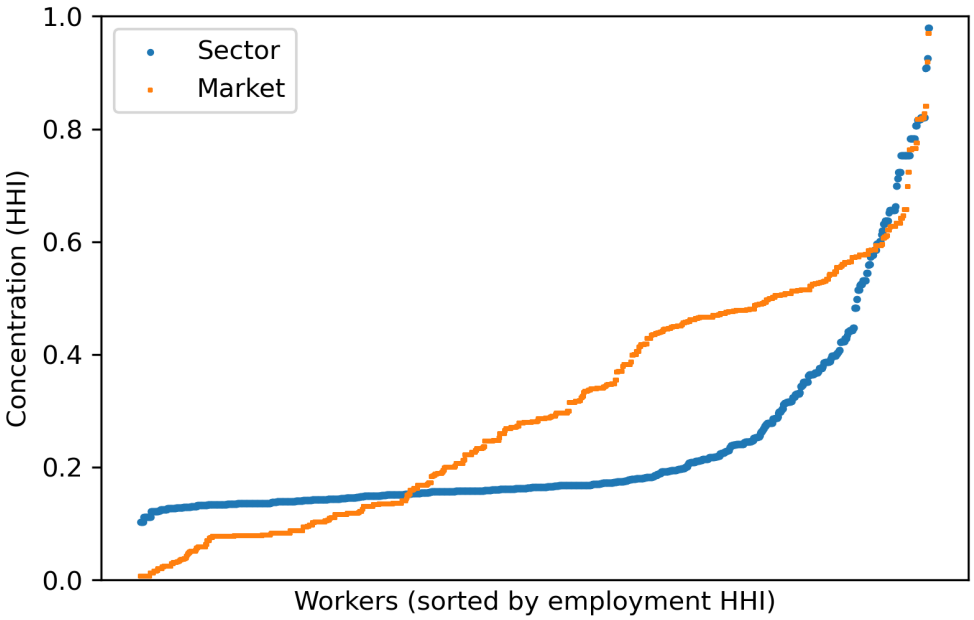
If our model is able to perfectly predict the actual effects of the Rio Olympics shock, the slope would be 1 and the intercept 0. As shown in the first column of Table 1.5 the slope of the best fit line is 0.982 and the intercept is -0.003, very close to our goals of 1 and 0, respectively.[26]

We further assess our model's predictive ability by comparing it to a series of standard approaches, which use our model but classify worker and job heterogeneity using commonly-used observable variables. Our first two standard approaches classify workers using 4-digit occupation codes instead of our network-based worker types. After dropping

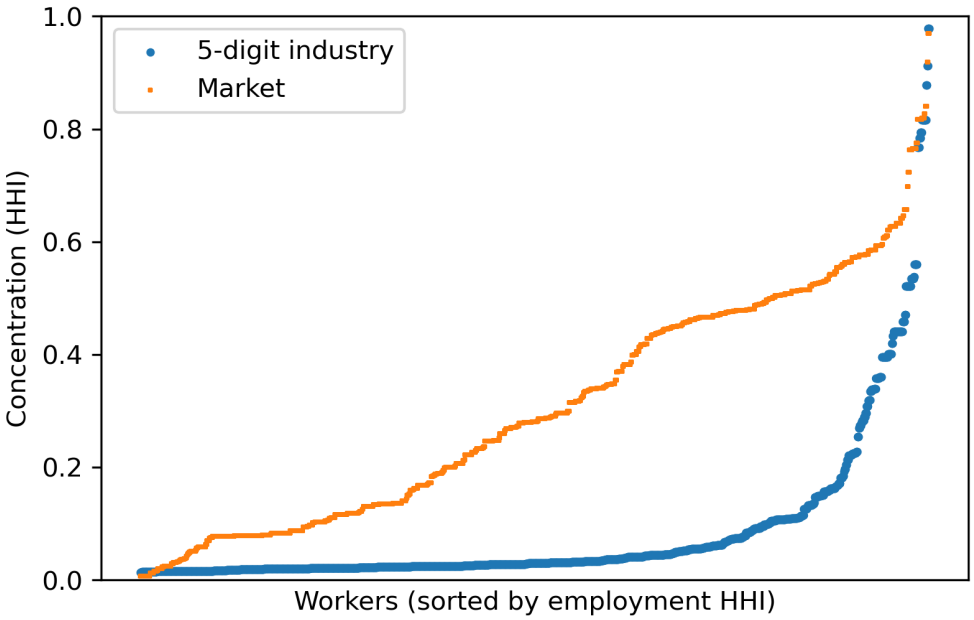
[26]The standard errors in this regression are large, but this is not surprising. There is significant variation that we are unable to predict because a number of important margins of adjustment are outside of our model. However, the fact that we estimate a slope close to 1 and an intercept close to 0 is consistent with these other factors being approximately orthogonal to our classifications. These other factors may include job amenities and non-monetary compensation, migration into or out of the Rio de Janeiro metro area, worker retraining, and changes in the tasks required by each job. Moreover, our model excludes linkages between sectors in the product market, which could affect demand for different types of labor, although our model could be expanded to include product market linkages by adding sector-level intermediate goods as inputs to firms' production functions (equation 1.2.3).

Figure 1.5: Concentration of Worker Types' ($\iota$) Employment Within Markets/Sectors

(a) Markets ($\gamma$) and sectors




(b) Markets ($\gamma$) and 5-digit industries




*Notes:* Figure presents concentration, defined as a Herfindahl-Hirschman Index (HHI), of worker types' employment within individual markets (orange lines) and sectors (blue line). The figure is weighted by the number of workers in each worker type. Workers are sorted from lowest to highest HHI along the horizontal axis. HHIs computed from the 2009-2012 RAIS worker earnings panel described in Section 1.5.

Table 1.5: Predicted Effect of Olympics on Wages: Network-Based vs. Standard Classifications

| Worker classification | $\iota$ | Occ4 | Occ4 | k-means | k-means |
|---|---|---|---|---|---|
| Job classification | $\gamma$ | sector | $\gamma$ | Sector | $\gamma$ |
| Intercept | -0.003 | -0.001 | -0.002 | -0.0 | -0.002 |
| | (0.009) | (0.009) | (0.009) | (0.011) | (0.01) |
| Model implied $\Delta$ log earnings | 0.982 | 0.148 | 0.428 | 0.234 | 0.566 |
| | (0.551) | (0.434) | (0.185) | (0.575) | (0.262) |
| MSE | 0.021 | 0.025 | 0.025 | 0.023 | 0.023 |
| Observations | 290 | 306 | 306 | 214 | 214 |

*Notes:* Table presents results from estimating equation (1.7.1) for various worker and job classifications. Workers classified by worker type ($\iota$) in column 1, 4-digit occupation in columns 2 and 3, and by k-means clusters of 6-digit occupations in columns 4 and 5. K-means clustering done on the basis of occupation specific skills defined by the U.S. O*NET, which is applied to Brazilian occupations using a crosswalk creaged by Aguinaldo Maciente (Maciente, 2013). Jobs are classified by market ($\gamma$) in columns 1, 3, and 5, and by IBGE sector in columns 2 and 4. Standard errors reported in parentheses. Independent and dependent variables defined at the worker classification level as described in Section 1.7. Dependent variables based on data from the 2009-2012 RAIS worker earnings panel described in Section 1.5. Independent computed by solving the model described in Section 1.2 using parameters estimated in Section 1.4.1 and calibrated in Section 1.4.2. Regressions are weighted by the number of workers per classification.

occupations with fewer than 5,000 employees for computational reasons,[27] we are left with 306 4-digit occupations, yielding a level of disaggregation similar to the 290 worker types. The second two benchmarks characterize worker heterogeneity using k-means clusters of 6-digit occupations based on 225 O*NET skills, where the number of clusters is chosen to match the number of worker types $\iota$, however we have to drop some of the resulting clusters because they are very small and are not observed in both the pre-shock and post-shock periods.[28] We classify job heterogeneity using sector in the first and third benchmark and using our network-based markets in the second and fourth. We present the results of these standard approaches in columns 2–5 of Table 1.5.

While our network-based classifications yield an approximately unbiased prediction of the actual shock-induced changes in earnings, the standard classifications do not. The coefficients on model-implied earnings changes are far below 1 in all four of the standard classifications. Moreover, the mean squared error (MSE) of our network-based classifications is below all four standard classifications. We interpret this as evidence in favor of our network-based

[27] This is necessary because we can only use occupations that are observed both pre-shock and post-shock.

[28] O*NET is defined for the U.S., but we use a crosswalk from the U.S. O*NET to the Brazilian occupation classification system created by Aguinaldo Maciente (Maciente, 2013). The clustering method yields a highly skewed cluster size distribution and we must drop some of the smallest clusters because they are not observed in both the pre-shock and post-shock periods. Therefore the actual number of clusters is somewhat smaller than the number of $\iota$'s.

classifications since they do a better job of predicting actual changes in the data than reasonable standard classifications.

## 1.8 Reduced form estimation of labor market shocks

A standard way of estimating the effects of labor demand shocks on workers is through the use of a Bartik instrument. A typical Bartik instrument measures the exposure of different groups of workers to labor demand shocks within groups of jobs. It can be written as

$$Bartik_g = \sum_s \pi_{gs} Shock_s \tag{1.8.1}$$

where $g$ defines a group of workers, $s$ defines a group of jobs, $\pi_{gs}$ is the fraction of group $g$ workers employed in group $s$ jobs before the shock, and $Shock_s$ is the size of the shock to group $s$ jobs. For example, in Autor et al.'s "China shock," $g$ represents commuting zones, $s$ indexes sectors, $\pi_{gs}$ is commuting zone $g$'s share of sector $s$ employment, and $Shock_s$ is the growth in Chinese imports in sector $s$. $Shock_s$ is a proxy for the size of the labor demand shock in sector $s$ jobs created by Chinese import growth, while $\pi_{gs}$ governs which workers are affected by the shock. Both $Shock_s$ and $\pi_{gs}$ depend upon the researcher's choice of classifications, $g$ and $s$, and therefore estimated effects of shocks are sensitive to these choices. In this section we study how the researcher's choice of worker and job classifications affect results.

We compare Bartik instruments based on our network-based worker types and markets to Bartik instruments based on occupations and sectors. First, we show that estimated effects of shocks on workers are significantly larger, as are $R^2$ values, when using our network-based classifications. Second, we provide a case study of a simulated shock in which we demonstrate that the reason why our worker types and markets yield larger coefficient estimates and $R^2$ values is that they more precisely identify which jobs experienced a change in demand for labor, and which workers were exposed to those jobs.

### 1.8.1 Analysis of the 2016 Rio de Janeiro Olympics

We begin by once again considering the labor demand shock created by the preparations for the Rio de Janeiro Olympics. As in Section 1.7, we define 2009 as the pre-shock period and 2014 as the post-shock period. We regress 2009 to 2014 changes in worker group $g$ earnings on the Bartik instrument defined in equation (1.8.1).

$$\Delta Earnings_g = \beta_0 + \beta_1 Bartik_g + \varepsilon_g \tag{1.8.2}$$

We have four specifications using all four combinations of our two worker classifications $g \in \{\text{worker type, occupation}\}$ and our two job classifications $s \in \{\text{market, sector}\}$. We normalize all of the Bartik instruments to have mean 0 and standard deviation 1 so that coefficients are directly comparable and can be interpreted as the effects of a 1 standard deviation change in the Bartik instrument on log earnings.[29] We measure $\pi_{gs}$ as the fraction of group $g$ workers who are employed in group $s$ jobs. $Shock_s$ is alternatively defined as the change in sector-level product output or changes in the market-level labor input, $\ell_\gamma$.

The results, presented in Table 1.6, show that estimated effects of the shock are highly sensitive to worker and job classifications. In column 1 we present our network-based classifications: workers are classified by worker type and jobs by market. In this specification, the effect of the shock on workers' earnings is positive and statistically significant, and the $R^2$ is large. The coefficient implies that a 1 standard deviation increase in exposure to the Olympics shock leads to an approximately 15.5% increase in earnings. Columns 2–4 present specifications using standard classifications. These specifications consistently find smaller (and in some cases negative) effects of the shock on workers, and have less explanatory power for variation in worker earnings, as shown by the smaller $R^2$ values. These results are consistent with occupation and sector doing a worse job of characterizing worker skill and job task heterogeneity than worker types and markets, and this misclassification leading to attenuated estimates and worse model fit.

While our results indicate that classifying worker and job heterogeneity with error yields attenuated estimates of effects *in this case*, it is not necessarily the case that classification errors of this sort yield estimates that are biased towards zero in general. Since we do not have classical measurement error, the intuition of measurement error leading to attenuation bias does not apply. In fact, there is no theoretical prediction about the direction of the bias due to misclasssification of workers and jobs in our context (Mahajan, 2006; Hu, 2008). We confirm this through a series of simulations in which we generate a data set according to the data generating process implied by our model, randomly misclassify varying percentages of workers and jobs, and then estimate the Bartik regression, equation (1.8.2). We find no clear relationship between the amount of misclassification and the slope coefficient $\hat{\beta}$. However, we do find that the $R^2$ values decline approximately monotonically with the fraction of workers and jobs misclassified. Therefore, we interpret the larger $R^2$ values from estimating equation (1.8.2) using our network-based classifications as evidence that the network-based classifications classify worker and job heterogeneity with less error than the standard classifications. By contrast, the larger coefficient estimate when we use our network-based

[29]Nonemployment is treated as 0 log earnings, so these regressions capture both movements in and out of employment and changes in earnings conditional on employment.

Table 1.6: Effects of exposure to Rio Olympics shock

| Exposure: | Market ($\gamma$) | Sector | Market ($\gamma$) | Sector |
|---|---|---|---|---|
| Worker classification: | Worker type ($\iota$) | Worker type ($\iota$) | Occ4 | Occ4 |
| Intercept | -0.169 | -0.169 | -0.156 | -0.156 |
| | (0.009) | (0.012) | (0.019) | (0.020) |
| iota exposure (market) | 0.156 | | | |
| | (0.009) | | | |
| iota exposure (sector) | | -0.031 | | |
| | | (0.012) | | |
| occ4 exposure (market) | | | 0.111 | |
| | | | (0.019) | |
| occ4 exposure (sector) | | | | -0.059 |
| | | | | (0.020) |
| Observations | 290 | 290 | 306 | 306 |
| $R^2$ | 0.531 | 0.021 | 0.096 | 0.027 |

*Note:*

*Notes:* Table presents the effect of the 2016 Rio de Janeiro Olympics shock on workers earnings from estimating equation (1.8.2). Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type ($\iota$) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data from the 2009-2012 RAIS worker earnings panel described in Section 1.5 aggregated to the worker classification level.

classifications is an empirical finding about the implications of misclassification in this context. See Appendix A.6 for details on these simulations.

Although the focus of this paper is classification of workers rather than identification of shocks, it is possible that the Olympics shock we study in this section may have been confounded by labor supply or other shocks. For example, workers may have anticipated the shock and migrated to Rio de Janeiro from other parts of Brazil. Therefore, in the next subsection we replicate the analysis in this subsection using simulated data in which we control the data generating process.

### 1.8.2 Reduced form analysis using simulated data

In this subsection, we demonstrate how estimated effects of shocks are sensitive to worker and job classifications in a setting where we can control the underlying data generating process. We replicate the analysis in the preceding section using simulated data. The simulated data have the same structure as the actual worker earnings panel described in

Section 1.5 that we used to estimate the labor supply parameters and for the empirical exercises in Sections 1.7 and 1.8.1, and are drawn from the data generating process defined by our model. Since we control the data generating process, we can be certain that we are observing an exogenous labor demand shock that is unconfounded by, for example, concurrent labor supply changes.

We generate the simulated data as follows. First, we calibrate demand shifters $\vec{a}^{Pre}$ to match the levels of product demand in each sector in 2009. We then solve the model using the 2009 demand shifters to generate a pre-shock wage vector $\vec{w}^{Pre}$ that clears all markets $\gamma$. We draw worker types and four-digit occupations from the empirical joint distribution of worker types and four-digit occupations. To generate job matches for each worker recall that, conditional on searching, workers choose a market to supply labor to according to equation (1.2.5):

$$\gamma_{it} = \underset{\gamma \in \{0,1,\ldots,\Gamma\}}{\arg\max}\ \psi_{\iota\gamma} w_{\gamma t} + \xi_{\gamma} + \varepsilon_{i\gamma t}.$$

This implies that a type $\iota$ worker chooses market $\gamma$ with probability given by equation (1.2.6):

$$\mathbb{P}_{\iota}[\gamma| = \frac{\exp\left(\frac{\hat{\psi}_{\iota\gamma} w_{\gamma t} + \hat{\xi}_{\gamma}}{\hat{\nu}}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\hat{\psi}_{\iota\gamma'} w_{\gamma' t} + \hat{\xi}_{\gamma'}}{\hat{\nu}}\right)},$$

where we use estimated parameter values $\hat{\Psi}$, $\hat{\Xi}$, and $\hat{\nu}$, estimated as described in Section 1.4. All workers make this choice in period $t = 1$, and in subsequent periods workers search again if they draw a separation shock as described in Assumption 1.2.2. In our full model, workers match with individual jobs after choosing markets, however the identity of the worker's individual job $j$ does not affect earnings or employment; it is only useful for classifying workers and jobs according to the BiSBM. Therefore, we do not specify the identity of each worker's specific job when generating our simulated data set.

Next, we draw sectors for each worker–job match according to the empirical joint distribution of sectors and markets. Finally, we draw earnings according to equation (1.4.1):

$$\omega_{it} = \psi_{\iota(i)\gamma_{it}} w_{\gamma_{it}} e_{it}.$$

where $e_{it}$ is log-normal measurement error. We repeat this exercise using the same labor supply parameters $\hat{\Psi}$, $\hat{\Xi}$, and $\hat{\nu}$ along with a new vector of demand shifters, $\vec{a}^{Post}$, calibrated to match the levels of product demand in each sector in 2014. We stack the two data sets to create a panel data set with both the pre-shock and post-shock periods.

We repeat the four Bartik-style regressions from the previous section using our simulated

Table 1.7: Effects of exposure to *simulated* Rio Olympics shock

| Exposure: | Market ($\gamma$) | Sector | Market ($\gamma$) | Sector |
|---|---|---|---|---|
| Worker classification: | Worker type ($\iota$) | Worker type ($\iota$) | Occ4 | Occ4 |
| Intercept | -0.139*** | -0.139*** | -0.147*** | -0.147*** |
| | (0.003) | (0.003) | (0.004) | (0.004) |
| iota exposure (market) | 0.018 | | | |
| | (0.003) | | | |
| iota exposure (sector) | | 0.014 | | |
| | | (0.003) | | |
| occ4 exposure (market) | | | 0.007 | |
| | | | (0.004) | |
| occ4 exposure (sector) | | | | 0.007 |
| | | | | (0.004) |
| Observations | 290 | 290 | 306 | 306 |
| $R^2$ | 0.090 | 0.053 | 0.009 | 0.011 |

*Note:*

*Notes:* Table presents the effect of the *simulated* 2016 Rio de Janeiro Olympics shock on workers earnings from estimating equation (1.8.2). Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type ($\iota$) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data generated using our model as the data generating process, as described in Section 1.8.2, and aggregated to the worker classification level.

data. The results, presented in Table 1.7, are qualitatively similar to the results using actual data in the previous section (Table 1.6), with the exception that the negative coefficients when jobs are classified by sector are now small positive coefficients. We continue to find larger coefficients and $R^2$ values when we define shock exposure according to markets as opposed to sectors, and when we classify workers according to worker type as opposed to 4-digit occupation. These results reiterate our point that misclassifying worker and jobs causes us to significantly understate the effects of shocks on workers in this context. In the next section we demonstrate that this is a more general finding.

### 1.8.3 Simulating many shocks

In the previous sections we found that the estimated effects of shocks are larger when using our network-based worker and job classifications than when using standard classifications. To allay any concern that our finding is specific to the Rio Olympics shock, we replicate the analysis in the previous section for a series of different shocks. For each of the 15 sectors,

Table 1.8: Means across all simulated shocks

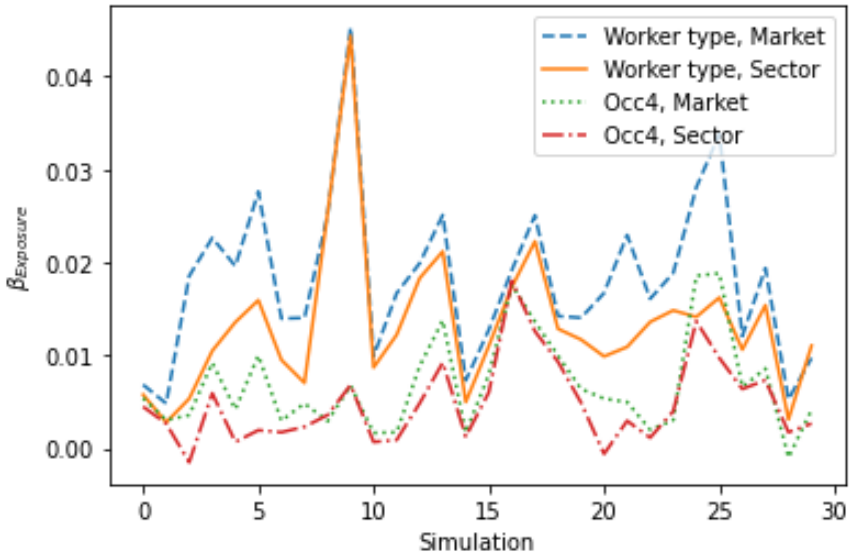
| Worker Classification | Job Classification | Coefficient | | $R^2$ | |
|---|---|---|---|---|---|
| | | Mean | Std Dev | Mean | Std Dev |
| Worker type | Market | 0.018 | 0.009 | 0.278 | 0.154 |
| Worker type | Sector | 0.013 | 0.008 | 0.167 | 0.147 |
| Occ4 | Market | 0.007 | 0.005 | 0.042 | 0.053 |
| Occ4 | Sector | 0.005 | 0.004 | 0.025 | 0.038 |

*Notes:* Table reports means and standard deviations of estimated regression coefficients and $R^2$ values from estimating the Bartik-style regression, equation (1.8.2), for each of the 30 simulated shocks described in Section 1.8.3. Workers classified by worker types ($\iota$) in rows 1 and 2, and by 4-digit occupation in rows 3 and 4. Jobs classified by market ($\gamma$) in rows 1 and 3, and by sector in rows 2 and 4.

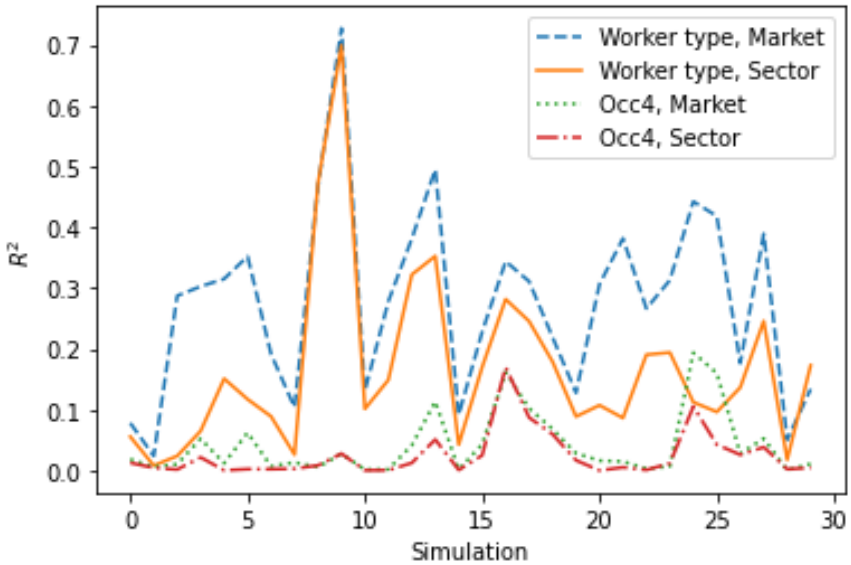
we simulate a positive shock in which the demand shifter for the shocked sector is doubled and the demand shifters for all other sectors are unchanged, and a negative shock in which the demand shifter for the shocked sector is halved and the demand shifters for all other sectors are unchanged. For each shock, we generate a new simulated data set and then use the simulated data to estimate the Bartik-style regression in equation (1.8.2) for each of the four combinations of worker and job classifications: $g \in \{\text{worker type, occupation}\}$ and $s \in \{\text{market, sector}\}$. We present the results in Table 1.8. We consistently find larger coefficients and $R^2$ values using our network-based classifications. The average coefficient from our network-based classification specification is 3.7 times larger than the average coefficient from the occupation–sector specification, and the average $R^2$ is 11 times larger. Figure 1.6 presents the slope coefficients and $R^2$ values from each individual regression in these simulations and shows that our network-based classifications yield slope coefficients and $R^2$ values that are uniformly larger than those from standard classifications, not just larger on average.

Figure 1.6: Exposure coefficients from all simulated shocks

(a) Slope coefficients




(b) $R^2$ values




*Notes:* Figure presents estimated regression coefficients and $R^2$ values from estimating the Bartik-style regression, equation (1.8.2), for each of the 30 simulated shocks described in Section 1.8.3.

### 1.8.4 Case study of shock to the "Accommodations and Food" sector

One of the shocks we simulated in the previous section was a 50% reduction in demand for the output of the Accommodations and Food sector, leaving the demand for all other sectors' output unchanged. This subsection explores that shock in greater detail to elucidate the mechanisms behind the finding that our network-based classifications yield larger estimates of the effects of shocks on workers. We focus on a shock to a single sector, as opposed to all sectors simultaneously as in the Rio Olympics shock, because this allows us to understand the precise nature of the shock.

Table 1.9 presents the same set of Bartik-style regressions as Tables 1.6 and 1.7 in the preceding sections. The qualitative story is unchanged: larger coefficients and $R^2$ values when we (i) define job heterogeneity according to markets as opposed to sectors, and (ii) when we define worker heterogeneity according to worker type as opposed to 4-digit occupation.

Table 1.9: Effects of exposure to simulated Accommodations and Food sector shock

| Job Classification: | Market ($\gamma$) | Sector | Market ($\gamma$) | Sector |
|---|---|---|---|---|
| Worker classification: | Worker type ($\iota$) | Worker type ($\iota$) | Occ4 | Occ4 |
| Intercept | -0.001 | -0.001 | -0.002 | -0.002 |
| | (0.001) | (0.001) | (0.002) | (0.002) |
| Effect of Shock | 0.007 | 0.006 | 0.001 | -0.001 |
| | (0.001) | (0.001) | (0.002) | (0.002) |
| Observations | 290 | 290 | 306 | 306 |
| $R^2$ | 0.070 | 0.047 | 0.001 | 0.000 |

*Note:*

*Notes:* Table presents the effect of the *simulated* Accommodations and Food sector shock on workers earnings from estimating equation (1.8.2). The shock is a 50% reduction in demand for the Accommodations and Food sector's output, holding demand for all other sectors' output constant. Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type ($\iota$) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data generated using our model as the data generating process, as described in Section 1.8.2, and aggregated to the worker classification level.

Why does the Bartik instrument have more explanatory power for workers' outcomes when workers are classified by worker types and jobs are classified by markets? On the worker side, it is because, as we argued in Sections 1.6.1 and 1.6.2, our worker types do a better job of identifying groups of homogenous workers than do occupations. We see this again by focusing on one of the worker types that was most affected by the shock to the Accommodations and Food sector, worker type $\iota = 64$. Table 1.10 tabulates the 10

Table 1.10: Occupation counts for $\iota$ =64

| Occ Code | Occ Description | Occ share |
|---|---|---|
| 513505 | Food services assistant | 0.090 |
| 521110 | Retail salesperson | 0.072 |
| 411005 | Office clerk | 0.043 |
| 514320 | Janitor | 0.032 |
| 513205 | General cook | 0.032 |
| 513215 | Industrial cook | 0.030 |
| 421125 | Cashier | 0.028 |
| 411010 | Administrative assistant | 0.026 |
| 763215 | Couturier, serial machining | 0.024 |
| 521125 | Stock clerk | 0.019 |

occupations we most frequently observe type $\iota = 64$ workers employed in. These occupations tend to be low-pay, low-education service sector occupations. The two most frequent are "food services assistant" and "retail salesperson." Our network-based classification method tells us that these retail and food services workers have similar skills despite the fact that they are employed in different occupations. If we had classified workers by occupation and jobs by sector, we would have implicitly assumed that the food services workers were exposed to the Accommodations and Food sector shock, while the retail salespeople were not. In reality, all of these workers were exposed to the shock because they have similar skills; workers not employed in the shocked sector may still be exposed to and affected by the shock if they are close substitutes for workers in the shocked sector. As we discussed in Section 1.8.1 and Appendix A.6, misclassifying workers such that some workers actually exposed to the shock are assumed not to have been exposed, and vice versa, leads to biased coefficient estimates and attenuated $R^2$ values.

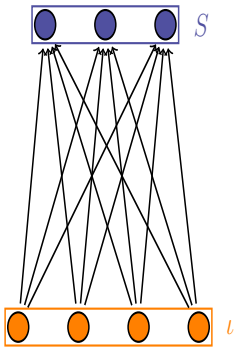
On the jobs side, classifying jobs by market rather than sector more accurately captures the channels through which shocks propagate from jobs to workers. Bartik instruments based on standard classifications assume that workers supply labor directly to sectors; our classifications allow workers to supply labor directly to markets but only indirectly to sectors, by way of markets (see Figure 1.7). We illustrate why our approach is prefereable by again focusing on type $\iota = 64$ workers. We have already established that these workers' skills are employable in both retail occupations and food service occupations, but who hires them? Do they supply labor to a retail market and a food services market? Or is there actually a market that includes jobs in both retail and food services? In Table 1.11 we present type $\iota = 64$ workers' labor supply by sector. Type $\iota = 64$ workers supply labor to a variety of sectors, including Retail, Wholesale and Vehicle Repair (28%) and Accommodations and

Food (14%). Since these workers supply labor to such a variety of sectors, no single sector can reasonably approximate the set of jobs to which they supply labor. By contrast, type $\iota = 64$ workers' labor supply *is* concentrated within specific network-based markets, $\gamma$.

Table 1.12 presents the percentage of their labor that type $\iota = 64$ workers supply to each market, restricting to the top 10. Type $\iota = 64$ workers supply over 60% of their labor to a single market, market $\gamma = 47$, and there is no other market to which they supply more than 3.5 percent of their labor. In other words, type $\iota = 64$ workers' labor supply is highly concentrated within a specific market, but not nearly as concentrated in specific sectors, despite the fact that we have vastly more markets (427) than sectors (15). This is a specific example of the more general finding of greater concentration of employment within markets than sectors that we presented in Section 1.6.3. Worker types' employment is more concentrated within markets than sectors because our markets are designed to identify groups of jobs that compete for similar workers, whereas sectors are defined by product markets. Therefore, our markets more closely approximate the channels through which shocks propagate through the labor market to workers. By contrast, classifying jobs by sectors introduces error by grouping together jobs with heterogeneous changes in labor demand. Again, as we discussed in Section 1.8.1 and Appendix A.6, misclassifying jobs such that jobs that in fact hire dissimilar workers are assumed to hire similar workers, and vice versa, leads to biased coefficient estimates and attenuated $R^2$ values.

Figure 1.7: Comparison of standard classifications and our model

(a) Standard Classifications




(b) Our Model

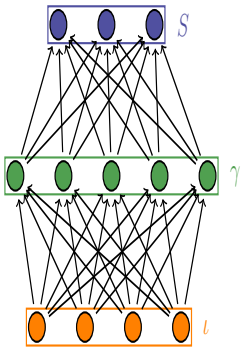


Table 1.11: Type $\iota = 64$ workers' labor supply by sector

| Sector | Share (%) |
|---|---|
| Retail, Wholesale and Vehicle Repair | 27.9 |
| Accommodation and food | 14.1 |
| Manufacturing industries | 11.7 |
| Professional, scientific and technical svcs | 11.1 |
| Arts, culture, sports and recreation and other... | 8.2 |
| Private health and education | 6.7 |
| Transport, storage and mail | 6.3 |
| Construction | 3.3 |
| Utilities | 2.6 |
| Extractive industries | 2.2 |
| Financial, insurance and related services | 2.2 |
| Information and communication | 2.1 |
| Public admin, defense, educ, health and soc se... | 1.4 |
| Real estate activities | 0.3 |
| Agriculture, livestock, forestry, fisheries an... | 0.1 |

*Notes:* Table presents the share of type $\iota = 64$ workers employed in each sector according to data generated by simulating the Accommodations and Food sector shock. The shock is a 50% reduction in demand for the Accommodations and Food sector's output, holding demand for all other sectors' output constant.

Table 1.12: Type $\iota = 64$ workers' labor supply by market ($\gamma$)

| Market ($\gamma$) | Share (%) |
|---|---|
| 47 | 60.2 |
| 189 | 3.5 |
| 116 | 1.7 |
| 242 | 1.5 |
| 418 | 1.3 |
| 83 | 1.3 |
| 36 | 1.2 |
| 138 | 1.1 |
| 125 | 0.9 |
| 45 | 0.8 |

*Notes:* Table presents the share of type $\iota = 64$ workers employed in each market ($\gamma$) according to data generated by simulating the Accommodations and Food sector shock. The shock is a 50% reduction in demand for the Accommodations and Food sector's output, holding demand for all other sectors' output constant. Only the 10 most frequently occurring markets are shown.

## 1.9 Conclusion

In this paper we develop a new method for clustering workers and jobs into discrete types that relies on workers' and jobs' choices, rather than observable variables or expert judgments. Our key insight is that linked employer-employee data contain a previously underutilized source of information: millions of worker–job matches, each of which reflects workers' and jobs' perceptions of the workers' skills and the jobs' tasks. We do so by microfounding a classification tool from the network theory literature with a Roy model of workers matching with jobs according to comparative advantage. The link between economic theory and network theory provides the worker types and markets we identify with a rigorous theoretical underpinning and clear interpretability.

We demonstrate that our network-based worker and job classifications outperform standard worker and job classifications in a number of ways. First, we show that an equilibrium model does a better job of predicting the effects of the Rio de Janeiro Olympics on workers' earnings when workers and jobs are classified using our network-based classifications than when they are classified using standard classifications. Second, we show that reduced form Bartik-style regressions yield larger and more precise estimates of the effects of shocks on workers when workers and jobs are classified using our network-based classifications as opposed to standard classifications.

A key feature of our classifications is that they simultaneously aggregate and disaggregate workers across occupations. They aggregate workers in different occupations who are revealed to have similar skills (for example, retail and food service workers), while disaggregating workers in the same occupation revealed to have distinct skills (for example course instructors focused on physical versus academic education). Our classifications, therefore, provide value beyond simply choosing the right granularity in, or aggregation of, occupation codes. They identify cohesive groups of workers and jobs that are not too granular to be useful in practical applications.

Although we apply our network-based clustering method to understanding the effects of labor market shocks on workers, this is only the beginning of our research agenda. We are currently working to apply different versions of the method to three different questions. First, we use our method to improve controls for worker skills in wage decompositions. Second, we use our worker and job classifications to improve measures of market power, based on the intuition that if retail and food services jobs compete for the same workers, they belong to the same market, even if they belong to different industries and occupations. Third, we are using closely related techniques to impute occupation and other worker characteristics in the LEHD.

Finally, although our current model abstracts from the role of physical space in the labor market and our empirics therefore focus on a single metropolitan area, we are working to expand our analysis to include geography and apply it to the entire country of Brazil. This will allow us to study the interaction of skills/tasks and geography in determining the scope of labor markets. For example, it will allow us to distinguish between different types of workers, likely with different types of skills, who search for jobs more nationally or more locally.

Our method is broadly applicable to important questions in labor economics and other fields. In addition to the applications to Bartik-style regressions we discuss in detail, our method may be useful any time researchers need to classify workers and/or jobs. For example, researchers studying how heterogeneous workers match with heterogeneous jobs might classify worker and job heterogeneity using our network-based classifications. The same is true for researchers studing the effects of shocks on workers using structural methods. More broadly, the method we develop may be used to classify agents using revealed preference any time agents' choices lead to a network structure of matches. For example, our method could be adapted to classify products and consumers based on detailed purchasing data, or to cluster financial institutions or countries based on networks of financial or trade flows. This paper provides a blueprint for doing so in a theoretically principled and data-driven way.

# CHAPTER II

# Building Better Counterfactuals for Gender Wage Gap Decompositions Using Matching and Network Theory (with Bernardo Modenesi)

## 2.1 Introduction

Significant attention has been paid to the gap in wages between men and women. Researchers are interested in understanding how much of the gap is due to men and women performing different work using different skills, and how much is due to men and women being paid differently for similar work. A number of methods exist for trying to answer this question. These methods decompose gender wage gaps into a portion explained by differences in characteristics between men and women, and a portion explained by differences in the return to characteristics, or "discrimination". However, all of these methods rely on three assumptions. First, they assume that unobserved determinants of earnings are independent of gender. To the extent that there exist unobserved worker characteristics that are important for determining wages and are correlated with gender, then researchers will obtain biased estimates of the return to observable characteristics. As a result, decompositions of gender wage gaps into a component explained by covariates and a component explained by the return to covariates will be incorrect. Second, they assume a functional form in order to estimate the function that maps observable characteristics into wages and thus serves as the foundation for counterfactuals that ask what men would earn if they had the same characteristics except their gender were switched to female, and vice versa. Third, they assume that the covariates for male workers and female workers share a common support. While this is likely to hold when the number of covariates is small, as more covariates are added (possibly to satisfy the independence assumption) the common support assumption

becomes more likely to be violated.[1]

In this paper, we (i) propose a new method for identifying unobserved determinants of workers' earnings from the information revealed by detailed data on worker–job matching patterns, (ii) non-parametrically estimate counterfactual wage functions for male and female workers, (iii) allow for a relaxation of the common support assumption, and (iv) apply our methods by decomposing the gender wage gap in Brazil using improved counterfactuals based on (i), (ii) and (iii). We find that the Brazilian gender wage gap is almost entirely explained by male and female workers who possess similar skills and perform similar tasks being paid different wages, not women possessing skills or tasks that pay relatively lower wages.

To understand the problem created by unobserved determinants of productivity, suppose that there are three types of worker characteristics that are relevant for determining wages: gender, other characteristics observable to researchers, and characteristics that are observable to labor market participants, but not to researchers. A naive wage decomposition would simply compare male wages to female wages and attribute all differences to the effect of gender. A more common approach would condition on observable characteristics like age, experience, occupation, education, and union membership and would attribute all differences in wages, conditional on these characteristics, solely to being a woman as opposed to being a man. However, this would miss the fact that even workers with identical observable covariates may perform distinct labor. As Goldin (2014) shows, male lawyers significantly outearn female lawyers largely because males are more likely to work long, inflexible hours, which leads to high wages. Therefore, if we simply compared the wages of male lawyers to the wages of female lawyers, we might mistakenly conclude that male and female lawyers receive differential pay for the same work, when in fact male and female lawyers perform different types of legal work. In other words, male and female lawyers differ in terms of covariates that are observed by labor market participants but not by researchers.

The key to our approach is identifying information about worker characteristics observable to labor market participants, but not to researchers, directly from the behavior of labor market participants. If we can identify groups of workers and groups of jobs who are similar *from the perspective of labor market participants*, then we can be confident that any gender wage differentials within these groups are due to differential returns to labor market activities by gender, rather than differences in the work done by male and female workers.

We employ a revealed preference approach that relies on workers' and jobs' choices, rather than observable variables or expert judgments, to classify workers and jobs into groups. Our key insight is that linked employer-employee data contain a previously underutilized source

[1] As more covariates are added it becomes harder to find another worker who shares the same values of all covariates.

of information: millions of worker–job matches, each of which reflects workers' and jobs' perceptions of the workers' skills and the jobs' tasks. Intuitively, if two workers are employed by the same job, they probably have similar skills, and if two jobs employ the same worker those jobs probably require workers to perform similar tasks. However, since discrimination may lead men and women with similar skills to sort into different jobs, our method includes a correction for gender-based sorting into jobs that normalizes workers' job match probabilities by the match probabilities for their gender.

We formalize this intuition and apply it to large-scale data using a Roy (1951) model in which workers supply labor to jobs according to comparative advantage. Workers belong to a discrete set of latent *worker types* defined by having the same "skills" and jobs belong to a discrete set of latent *markets* defined by requiring employees to perform the same "tasks."[2] Workers match with jobs according to comparative advantage, which is determined by complementarities between skills and tasks at the worker type–market level. Workers who have similar vectors of match probabilities over markets are therefore revealed to have similar skills and belong to the same worker type, and jobs that have similar vectors of match probabilities over worker types are revealed to have similar tasks and belong to the same market. Our model extends the model in Fogel and Modenesi (2022) to allow firms to have labor market power, thereby rationalizing pay heterogeneity among workers with the same skills in jobs requiring the same tasks and microfounding the correction for gender-based sorting.

Once we have clustered workers with similar skills into worker types and jobs requiring similar tasks into markets, we turn to estimating counterfactual wage functions. Mainstream decomposition methods estimate counterfactual female earnings by fitting wage regressions using observations for male workers only, but generating predicted values by multiplying average female covariate values by the male regression coefficients. This approach suffers from three main issues: (i) it requires the researcher to impose a restrictive regression functional form; (ii) it does not necessarily allow for heterogeneous returns to covariates in predictions; and (iii) it does not have embedded tools to handle when workers do not share similar covariate support. Taken together, these issues can potentially bias the counterfactual estimation exercise, which is the foundation of gender wage gap decompositions. In order to circumvent these issues, we make use of a flexible matching estimator for counterfactual earnings.

We implement a matching estimator in which we match male and female workers who belong to the same worker type and are employed by jobs in the same market. In doing

[2] "Skills" and "tasks" should be interpreted broadly as any worker and job characteristics that determine which workers match with which jobs.

so, we implicitly assume that worker types and markets account fully for all factors, other than gender, that affect workers' wages, although we also estimate specifications in which we include other observable characteristics in addition to worker types and markets. Within these matched groups, we use the male workers' mean wages as counterfactuals for what the female workers would have earned if they were male, and vice versa. We compare our matching estimator to a standard estimator and find similar results, although in some specifications the matching estimator is clearly preferable. However, there may be some worker type–market cells with no male workers or no female workers so we introduce a correction to account for this lack of common support.

We address the issue of a lack of common covariate support between male and female workers by decomposing the gender wage gap into four components: (i) differences due to different covariate distributions between groups, i.e. the *composition factor*, for observations that share the same support; (ii) differences related to differential returns to covariates between groups over a common support of the covariates, i.e. the *structural factor*, often associated with labor market discrimination; (iii) a part due to observations from male workers being out of the female workers' support of the covariates; and (iv) the last portion related to observations of female workers being out of the male workers' support of the covariates. This decomposition allows us to perform counterfactuals similar to existing methods for the part of the distribution of the covariates for which male and female workers have common support, yet it still allows us to quantify how much of the gender wage gap occurs outside the region of common support and would therefore be ignored by standard decomposition methods.

We estimate our model and conduct empirical analyses using Brazilian administrative records from the Annual Social Information Survey (RAIS) that is managed by the Brazilian labor ministry. The RAIS data contain detailed information about every formal sector employment contract, including worker demographic information, occupation, sector, and earnings. Critically, these data represent a network of worker–job matches in which workers are connected to every job they have ever held, allowing us to identify job histories of workers, their coworkers, their coworkers' coworkers, and so on. We restrict our analysis to the Rio de Janeiro metropolitan area both for computational reasons and because restricting to a single metropolitan area enables us to focus on skills and tasks dimensions of worker and job heterogeneity rather than geographic heterogeneity.

In our data, the average male worker earns a wage 16.7% higher than the average female worker. Our primary result is that almost the entire gender wage gap is attributable to male and female workers who possess similar skills and perform similar tasks being paid differently, or what is often referred to as "discrimination." This is true at the aggregate level, and

remains true when we perform wage decompositions within each worker type–market cell, indicating that this is a widespread phenomenon, not one driven by large wage differentials in small subsets of the labor market. We find that wage decompositions based on standard observable variables suffer from omitted variable bias, emphasizing the need for detailed worker and job characteristics in the form of worker types and markets. We find that wage decompositions based on linear regressions yield similar findings to those based on matching when a lack of common support is not an issue, however when male and female workers' characteristics do not share a common support the matching estimator with corrections for a lack of common support outperforms alternatives.

**Literature:** The literature of decomposition methods in economics can be classified into two main branches. The first decomposes average differences in a variable of interest $Y$ — often wages — between two groups of workers. The most widespread method in this class was developed by Oaxaca (1973) and Blinder (1973). The second branch decomposes functionals of the variable of interest $Y$ – e.g. its distribution or quantile function. Given that functionals of a variable often provide more information than its average, the second group of decompositions is referred to as "detailed decompositions" (Fortin et al., 2011). A seminal paper in this group is DiNardo et al. (1996)[3] and their methodology and inference was further generalized and improved later by Chernozhukov et al. (2013)[4]. We follow the first branch of the literature in focusing on average differences, largely because our rich set of controls introduces a curse of dimensionality that renders detailed decompositions infeasible.

Our method for handling a lack of common covariate support follows Nopo (2008) and Garcia et al. (2009)[5]. In concurrent work we extend Nopo (2008) to generic "detailed decompositions" (Modenesi, 2022).

Our model of labor market power builds on Card et al. (2015), Card et al. (2018) and Gerard et al. (2018) but allows for significantly more granular worker and job heterogeneity. The way we model multidimensional worker–job heterogeneity relates to papers that use a skills-tasks framework in the worker-job matching literature (Autor et al., 2003; Acemoglu and Autor, 2011; Autor, 2013; Lindenlaub, 2017; Tan, 2018; Kantenga, 2018). Our method

[3]Barsky et al. (2002) develop a methodology similar to DiNardo et al. (1996), focusing on issues that arise from lack of common covariate support between the groups in the decomposition. Modenesi (2022) discusses their approach in light of alternatives to handle the lack of common support.

[4]Firpo et al. (2018) later in this literature uses influence functions to propose a detailed decomposition that is invariant to the order of the decomposition.

[5]Garcia et al. (2009) and Morello and Anjolim (2021) both study the evolution of the Brazilian gender gap. Garcia et al. (2009) uses the same approach we use to handle the problem of lack of overlapping supports, and Morello and Anjolim (2021) have a similar matching methodology to decompose the gender gap. In addition to using similar methods for the decomposition, we add the skills and tasks controls derived from the labor market network, and we derive a distribution of gender gaps for different clusters of similar workers performing similar tasks.

for clustering workers and jobs fits into the relatively recent literature in labor economics that extracts latent information from the network structure of the labor market (Sorkin, 2018; Nimczik, 2018; Jarosch et al., 2019) and directly extends Fogel and Modenesi (2022) by allowing for labor market power. Methodologically, we draw from the community detection branch of network theory (Larremore et al., 2014; Peixoto, 2018; 2019)[6].

By controlling for skills and tasks, our papers share common ground with Goldin (2014) and Hurst et al. (2021). Goldin (2014) indicates that the potential residual discrimination in the gender wage gap is due to the nature of the tasks in some occupations, by using a linear regression approach dummies for occupation interacted with the gender dummy. We add to her approach by proposing an economic model for discrimination, which provides us with *both* worker and job heterogeneity controls, in addition to performing the gender gap decomposition while taking into account potential violations of conventional decomposition assumptions. Hurst et al. (2021) on the other hand are assessing the black-white wage gap over time as function of changes in the taste vs statistical discrimination factors, as well as the result of workers sorting after these changes.

**Roadmap:** The paper proceeds as follows. Section 2.2 introduces a simple framework for decomposition methods. Section 2.3 presents our model of worker–job matching and derives from it our algorithm for clustering workers into worker types and jobs into markets. Section 2.4 provides greater detail on the wage gap decomposition methods we employ. Section 2.5 describes our data. Section 2.6 presents results. Finally, Section 2.7 concludes.

## 2.2 A framework for decomposition methods

We introduce a simple framework for decomposition methods to guide the analysis in this paper. Define the actual wage of worker $i$ employed by job $j$ as $Y_{ij}$, and let $G_i$ be a dummy denoting whether worker $i$ is male. The difference between the average wage for male workers and the average wage for female workers, which we call the "overall wage gap," can be expressed as:

$$\Delta := E[Y_{ij}|G_i = 1] - E[Y_{ij}|G_i = 0] \tag{2.2.1}$$

The overall wage gap above can be decomposed into two factors: differences in productivity between male and female workers, usually referred to as the *composition* factor;

[6]More precisely, we employ a variant of the SBM which makes use of network edge weights (Peixoto, 2018), which are key for us to model the presence of potential discrimination in the labor market. Our paper connects to this literature by formalizing a theoretical link between monopsonistic labor market models and the stochastic block model, providing microfoundations and economic interpretability of network theory unsupervised learning tools in order to solve economic problems.

and differences in pay between equally productive male and female workers, known as the *structure* factor. We use the potential outcomes framework in order to formally decompose the overall wage gap into these two factors. Denote by $Y_{0ij}$ the potential wage of worker $i$ employed by job $j$ when the worker is female, and $Y_{1ij}$ the potential wage of worker $i$ employed by job $j$ when the worker is male. Let $x$ be the vector of all variables that determine workers' productivity. We assume that the worker's gender may affect their pay, but does not directly affect their productivity. We represent the potential outcomes as functions of $x$ as follows: $Y_{gij} := Y_g(x_{ij}), g \in \{0, 1\}$. Notice that $x$ has both $i$ and $j$ subscripts, as the marginal product of worker $i$ at their current job $j$ depends on both the worker's skills and the job's tasks. The fact that there is a different earnings function for men and women reflects the possibility that male and female workers with identical productivities may be paid differently. Furthermore, it is possible to use the dummy for gender to represent observed wages as a function of potential outcomes using a switching regression model $Y_{ij} := G_i Y_g(x_{ij}) - (1 - G_i)Y_g(x_{ij})$.

At this point we are able to decompose the overall wage gap, $\Delta$, into the *composition* and *structure* components mentioned above by adding and subtracting the quantity[7]

$$E[Y_1(x_{ij})|G_i = 0] := \int Y_1(x_{ij}) dF_{G=0}(x)$$

from the overall wage gap $\Delta$, where $F_{G=0}(x)$ is the productivity distribution for female workers. Intuitively, $E[Y_1(x_{ij})|G_i = 0]$ is the mean earnings for a counterfactual set of workers possessing the female productivity distribution, but who are paid like men[8].

$$\Delta := \underbrace{E[Y_{ij}|G_i = 1] - E[Y_1(x_{ij})|G_i = 0]}_{\Delta_X := \text{Composition}} + \underbrace{E[Y_1(x_{ij})|G_i = 0] - E[Y_{ij}|G_i = 0]}_{\Delta_0 := \text{Structure}} \tag{2.2.2}$$

The *composition* portion can be rewritten as $E[Y_1(x_{ij})|G_i = 1] - E[Y_1(x_{ij})|G_i = 0]$[9]. It represents the difference between what male workers actually earn and what male workers would have earned in a counterfactual scenario in which their productivity distribution was equivalent to the female productivity distribution. This quantity captures the portion of the overall wage gap attributable to differences in the composition, or distribution of productivity, between male and female workers. The *structure* portion is equivalent

[7]Analogously, the overall decomposition can be performed by adding and subtracting the male counterfactual quantity $E[Y_0(x_{ij})|G_i = 1]$ to $\Delta$. The main results in this paper use the female counterfactual approach.

[8]Alternatively, this counterfactual term can be interpreted as the mean earnings of male workers whose productivity distribution was adjusted to match the female productivity distribution

[9]We use the representation of the observed $Y$ in terms of potential outcomes to write $E[Y_{ij}|G_i = 1] = E[G_i Y_g(x_{ij}) - (1 - G_i)Y_g(x_{ij})|G_i = 1] = E[Y_1(x_{ij})|G_i = 1]$ and substitute it in $\Delta_X$.

to $E[Y_1(x_{ij}) - Y_0(x_{ij})|G_i = 0]$[10]. This is the difference between female earnings in a counterfactual state in which females were paid equivalently to what equally productive male workers are paid and actual average female earnings. This portion of the overall wage gap is due to structural differences in how the two genders are paid, holding productivity constant, which is why this term is often associated with a form of discrimination.

What we define as the structural component might reasonably be thought as discrimination, where labor market discrimination is defined as workers with similar productivity, performing similar tasks, and being paid differently based on observables that do not influence productivity. Other forms of discrimination may exist — including mistreatment or harassment, differential pre-job human capital accumulation opportunities, or discriminatory hiring practices — but we do not consider those in this paper. In our set up, individual discrimination occurs when the wage for worker $i$ at job $j$ is different if the individual's gender changes, *ceteris paribus*, i.e. $Y_1(x_{ij}) - Y_0(x_{ij}) \neq 0$. The problem is that, in order to measure this quantity, we run into the fundamental problem of causal inference: it is impossible to observe the potential wages in both states for the same individual. Therefore we must make assumptions in order to construct counterfactual values, i.e. the value of $Y_1$ for a female worker, or the value of $Y_0$ for a male worker. In this paper, we break the assumptions needed for the counterfactual estimation into two parts and we show how our approach contributes to deal with limitations in each of them.

The first assumption is that workers with the same values of $x$ are equally productive and would be paid equal wages if gender played no role in wage determination, conditional on productivity. This is equivalent to assuming that $x$ contains all factors that affect productivity and are correlated with gender. This "conditional independence/ignorability" assumption, is the basis of all decomposition methods in economics (Fortin et al., 2011), as it is a requirement for consistency of its estimates for the gap decomposition portions. However, not all factors that theoretically should be included in $x$ are observable.

A problem would arise if certain factors that contribute to worker $i$'s productivity in job $j$ are both unobserved by the econometrician and correlated with gender. If such factors exist, our counterfactuals would be invalid. Specifically, wage differentials due to unobserved differences in skills and tasks between male and female workers would be attributed to the effect of gender itself. For example, if women tend to have better social skills but we do not observe social skills, then we would interpret women outearning men in social skill-intensive jobs as discrimination against men, when in fact it is simply the result of differences in

[10] Analogously to the previous term, using the map from the potential outcomes to the observed $Y$, we can write $E[Y_{ij}|G_i = 0] = E[G_i Y_g(x_{ij}) - (1 - G_i)Y_g(x_{ij})|G_i = 1] = E[Y_0(x_{ij})|G_i = 0]$ and substitute it in $\Delta_0$.

unobserved skills. Therefore, it is critical to come as close as possible to identifying groups of male and female workers who have exactly the same skills and perform exactly the same tasks. If we do so, then any gender wage differentials within this group are attributable to the effect of gender *per se*. In Section 2.3 we address this issue by identifying latent worker and job characteristics relevant to productivity and wage determination using the network of worker–job matches.

The second set of assumptions required to build the counterfactual $Y_1(x)$ for females in $\Delta$ are related to the choice of an estimation strategy for the function $Y_1(\cdot)$[11]. A common estimation strategy requires fitting a linear wage regression for males and using its estimated coefficients to predict wages, but inputting female workers' covariates (Oaxaca, 1973 and Blinder, 1973). This approach is highly tractable, however the assumption of a linear functional form is to some extent arbitrary, and using the same regression coefficients to predict counterfactual earnings for distinct female workers (i.e. allowing no heterogeneous returns to observable characteristics) could lead to biased estimates of counterfactual earnings. An alternative approach relies on matching males to each female worker based on similar observable characteristics, and uses the wages of matched male workers in order to inform each female's counterfactual wage. This less-parametric approach has the advantage of not imposing any functional form assumption for $Y_1(\cdot)$, however it requires us to observe a sufficiently rich set of observable variables that male and female workers with the same observables may be assumed to have similar productivity. Moreover, matching methods are unreliable when we are unable to find a female worker with the same observables as a male worker, or vice versa. In Section 2.3 we describe a new method to enhance the set of observable characteristics available to the researcher, reducing the scope for unobserved determinants of productivity to cause biased estimates. In Section 2.4 we compare and contrast different methods to decompose the gender wage gap given a set of observable characteristics, circumventing issues present in counterfactual earnings estimation.

[11] Another approach decomposes the wage *distributions*, as opposed to actual wages, which would be equivalent to switching $Y$ for its distribution $F_Y$, but still needing the estimation of the counterfactual $F_{Y_1}(y|x)$ for females (e.g. DiNardo et al., 1996 and Chernozhukov et al., 2013). We choose not to employ these decompositions in this paper as our setup does not satisfy basic conditions for decomposing distributions, such as having a low-dimensional vector of observable characteristics $x$ – given curse of dimensionality – and having the overlapping supports assumptions satisfied.

## 2.3 Revealing latent worker and job heterogeneity using network theory

In this section we present an economic model of monopsonistic wage setting, which rationalizes a wage gap between two groups of workers who have different demographic characteristics, but have the same skills and perform the same tasks. Intuitively, otherwise identical male and female workers may supply labor to individual jobs with different elasticities, and jobs respond by offering them wages with different markdowns. If one group of workers supply labor to jobs more inelastically, then they will be paid less, holding productivity constant. Moreover, the model microfounds our network-based clustering algorithm, which identifies groups of male and female workers with similar skills who perform similar tasks, and therefore can serve as good counterfactuals for each other. The model builds on the model of the labor market developed in Fogel and Modenesi (2022), with two important differences: (i) in this paper workers have idiosyncratic preferences over individual jobs, not just markets, causing jobs to face upward-sloping labor supply curves, and (ii) firms may offer different wages to men and women, even if they have identical skills and perform identical tasks. The model defines a probability distribution that governs how workers match with jobs, forming the network of worker-job matches observed in linked employer-employee data. We use this probability distribution to assign similar workers to worker types and similar jobs to markets, using a Bayesian method based on generative network theory models, which we present after the economic model.

### 2.3.1 Economic model

We propose a model with two primary components: heterogeneous workers who supply labor and firms that produce goods by employing labor to perform tasks. Workers supply their skills to jobs, which are bundles of tasks embedded within firms. Jobs' tasks are combined by the firms' production functions to produce output. We assume that firms face an exogenously-determined demand for their goods.[12] Our model of the labor market has the following components:

- Each worker is endowed with a "worker type," and all workers of the same type have the same skills.
- A job is a bundle of tasks within a firm. As we discuss in Section 2.5, we define a job in our data as an occupation–establishment pair.

[12]For an alternative version of the model with endogenous product demand, see Fogel and Modenesi (2022).

- Each job belongs to a "market," and all jobs in the same market are composed of the same bundle of tasks.
- There are $I$ worker types, indexed by $\iota$, and $\Gamma$ markets, indexed by $\gamma$.
- The key parameter governing worker-job match propensity is an $I \times \Gamma$ productivity matrix, $\Psi$, where the $(\iota, \gamma)$ cell, $\psi_{\iota\gamma}$ denotes the number of efficiency units of labor a type $\iota$ worker can supply to a job in market $\gamma$.[13]

Time is discrete, with time periods indexed by $t \in \{1, \ldots, T\}$ and workers make idiosyncratic moves between jobs over time. Neither workers, households, nor firms make dynamic decisions, meaning that the model may be considered one period at a time. We do not consider capital as an input to production.

#### 2.3.1.1 Firm's problem

Each firm, indexed by $f$, has a production function $Y_f(\cdot)$ which aggregates tasks from different labor markets, indexed by $\gamma$. Firm $f$ faces exogenously-determined demand for its output, $\bar{Y}_f$. The firm's only cost is labor. As we discuss in the next subsection, firms face upward-sloping labor supply curves and therefore have wage-setting power. Firms demand labor in each market, $\gamma \in \{1, \ldots, \Gamma\}$ and offer a different wage per efficiency unit of labor for each market. Firms also may offer different wages to workers in different demographic groups $g \in \{A, B\}$ (e.g. male and female workers), although type $A$ and type $B$ workers belonging to the same worker type $\iota$ are equally productive in all jobs. We define a job $j$ as a firm $f$ – market $\gamma$ pair. We define the wage per efficiency unit of labor for demographic group $g$ workers employed in job $j$ $w_j^g$. Define $L_j^g$ as the quantity of efficiency units of labor supplied by demographic group $g$ workers to job $j$.

The firm's problem is to choose the quantity of labor inputs in each job for each demographic group in order to minimize costs subject to the constraint that production is greater than or equal to the firm's exogenous product demand, $\bar{Y}_f$:

$$\min_{\{w_j^A, w_j^B\}_{j=1}^{\Gamma}} \sum_{j=1}^{\Gamma} w_j^A L_j^A + w_j^B L_j^B \quad \text{s.t.} \quad Y_f\left(L_1, \ldots, L_\Gamma\right) \geq \bar{Y}_f$$

[13] We can think of $\psi_{\iota\gamma}$ as $\psi_{\iota\gamma} = f(X_\iota, Y_\gamma)$, where $X_\iota$ is an arbitrarily high dimensional vector of skills for type $\iota$ workers, $Y_\gamma$ is an arbitrarily high dimensional vector of tasks for jobs in market $\gamma$, and $f()$ is a function mapping skills and tasks into productivity. This framework is consistent with Acemoglu and Autor (2011)'s skill and task-based model, and is equivalent to Lindenlaub (2017) and Tan (2018). A key difference is that Lindenlaub and Tan observe $X$ and $Y$ directly and assume a functional form for $f()$, whereas we assume that $X$, $Y$, and $f()$ exist but are latent. We do not identify $X$, $Y$, and $f()$ directly because in our framework $\psi_{\iota\gamma}$ is a sufficient statistic for all of them.

where $L_j = L_j^A + L_j^B$ is the total amount of efficiency units of labor employed by job $j$ and $Y_f$ is a concave and differentiable production function.

Taking the first order condition with respect to $w_j^g$ allows us to solve for the wage paid by job $j$ to workers in demographic group $g$ as a markdown relative to the marginal revenue product of labor:

$$w_j^g = \underbrace{\frac{e_j^g}{1+e_j^g}}_{\text{Markdown}} \times \underbrace{\mu_f \frac{\partial Y_f}{\partial L_j}}_{\text{Marg. revenue product of labor}} \tag{2.3.1}$$

where $\mu_f$ is the shadow revenue associated with one more unit of output and $e_j^g := \frac{\partial L_j^g}{\partial w_j^g} \frac{w_j^g}{L_j^g}$ is the labor supply elasticity of workers from group $g$ to job $j$.

Equation (2.3.1) shows that the wage paid to demographic group $g$ workers employed in job $j$ (equivalently, employed in market $\gamma$ by firm $f$) is the product of a markdown and the marginal revenue product of labor in job $j$. The markdown depends on demographic group $g$'s elasticity of labor supply to job $j$. As labor supply becomes more elastic, the markdown converges to 1 and the wage converges to the marginal product of labor. Conversely, as labor supply becomes less elastic, the wage declines further below the marginal product of labor. This equation rationalizes different demographic groups being paid different wages for the same labor: if one demographic group supplies labor more inelastically, they will be paid less.[14] The firm employs workers in both demographic groups despite paying them different wages because in order to attract the marginal worker from the lower-paid demographic group, it must raise wages for all inframarginal workers in that group. At some point the marginal cost (inclusive of the required raises for inframarginal workers) of hiring workers from the lower-paid demographic group exceeds the marginal cost of hiring workers from the higher-paid demographic group, and the firm will switch to hiring the higher-paid workers.

#### 2.3.1.2 Worker's problem

A worker belonging to worker type $\iota$ and demographic group $g \in \{A, B\}$, has a two step decision. First, she chooses a market $\gamma$ in which to look for a job, and second she chooses a firm $f$ (and by extension a job $j$). The worker's type defines their skills. Type $\iota$ workers can supply $\psi_{\iota\gamma}$ efficiency units of labor to any job in market $\gamma$. $\psi_{\iota\gamma}$ is a reduced form representation of the skill level of a type $\iota$ worker in the various tasks required by a

[14] We are referring to the elasticity of labor supply *to a specific job* $j$, which may differ from a group's labor supply elasticity to the overall labor market. For example, it could be the case that men supply labor more inelastically at the extensive margin, but women have stronger idiosyncratic preferences for specific jobs, making them less likely to change jobs in response to a wage differential. In this case, women would supply labor less elastically to a specific job $j$ and thus receive lower wages.

job in market $\gamma$. Units of human capital are perfectly substitutable, meaning that if type 1 workers are twice as productive as type 2 workers in a particular market $\gamma$ (i.e. $\psi_{1\gamma} = 2\psi_{2\gamma}$), firms would be indifferent between hiring one type 1 worker and two type 2 workers at a given wage per efficiency unit of labor, $w_j$. Therefore, the law of one price holds within each demographic group for each job, and a type $\iota$ worker belonging to demographic group $g$ employed in a job in market $\gamma$ is paid $\psi_{\iota\gamma} w_j^g$. Because workers' time is indivisible, each worker may supply labor to only one job in each period and we do not consider the hours margin.

Workers choose job $j$, equivalent to $\gamma f$, in order to maximize utility, which is the sum of log earnings $\log(\psi_{\iota\gamma} w_j^g)$ and an idiosyncratic preference for job $j$, $\varepsilon_{ij}^g$:

$$j^* = \arg\max_j \quad \log(\psi_{\iota\gamma} w_j^g) + \varepsilon_{ij}^g.$$

We assume that $\varepsilon_{ij}^g$ follows a nested logit distribution with parameters $\theta^g$ and $\nu_{\iota\gamma}^g$, with the $\iota\gamma$ subscript on the latter parameter indicating that nests are defined as $\iota, \gamma$ pairs:

$$\varepsilon_{ij}^g \sim NestedLogit(\theta^g, \nu_{\iota\gamma}^g)$$

It follows from this assumption about the distribution of $\varepsilon_{ij}^g$ that the probability that worker $i$ belonging to worker type $\iota$ and demographic group $g$ matches with job $j$ in market $\gamma$ is:

$$P(j = j^* | j \in \gamma, i \in \iota, g) = \underbrace{\underbrace{\frac{(I_{\iota\gamma}^g)^{\frac{1}{\nu_{\iota\gamma}^g}}}{\sum_\gamma (I_{\iota\gamma}^g)^{\frac{1}{\nu_{\iota\gamma}^g}}}}_{P(\gamma=\gamma^* | i\in\iota, j\in\gamma, g)}}_{\textbf{1st step}\text{: market choice}} \underbrace{\underbrace{\frac{(\psi_{\iota\gamma} w_j^g)^{\frac{1}{\theta^g}}}{\sum_{j\in\gamma} (\psi_{\iota\gamma} w_j^g)^{\frac{1}{\theta^g}}}}_{P(j=j^* | i\in\iota, j\in\gamma, \gamma=\gamma^*, g)}}_{\textbf{2nd step}\text{: job choice}} \tag{2.3.2}$$

where $I_{\iota\gamma}^g := \sum_{j\in\gamma} (\psi_{\iota\gamma} w_j^g)^{\frac{1}{\theta^g}}$, also referred to as the inclusive value, is the expected utility a type $\iota$ worker faces when choosing market $\gamma$. Intuitively, the nested logit assumption decomposes the job choice probability into a first stage in which the worker chooses a market and then a second stage in which the worker chooses a job conditional on their choice of a market.

### 2.3.2 Identifying worker types and markets

#### 2.3.2.1 Deriving the likelihood

Now that we have derived the probability of worker $i$ matching with job $j$ from the primitives of our model, the next step is using this probability as the basis for a maximum

likelihood procedure that assigns workers to worker types and jobs to markets based on the observed set of worker–job matches. This procedure builds on Fogel and Modenesi (2022), by allowing workers in the same worker type but different demographic groups to have different vectors of match probabilities over jobs.

We decompose the choice probability in equation (2.3.2) into a component that depends only on variation at the $\iota, \gamma, g$ level and a component that depends on wages at individual jobs:

$$P(j = j^* | j \in \gamma, i \in \iota, g) = \underbrace{\frac{(I^g_{\iota\gamma})^{\frac{1}{\nu^g_{\iota\gamma}}-1}}{\sum_\gamma (I^g_{\iota\gamma})^{\frac{1}{\nu^g_{\iota\gamma}}}} \psi^{\frac{1}{\theta^g}}_{\iota\gamma}}_{\equiv \Omega^g_{\iota\gamma}} \underbrace{(w^g_j)^{\frac{1}{\theta^g}}}_{\equiv d^g_j} \quad . \tag{2.3.3}$$

$\iota-\gamma-g$ component; $j-g$ component

The first term reflects workers choosing markets according to comparative advantage, while the second captures the fact that some jobs in market $\gamma$ require more workers than others (due to exogenous product demand differences), and since jobs face upward-sloping labor supply curves, they must pay higher wages to attract greater numbers of workers. Isolating the group-level $(\iota, \gamma, g)$ variation from the idiosyncratic job-level variation allows us to cluster workers into worker types and jobs into markets on the basis of having the same group-level match probabilities, as we discuss below.

The choice probabilities we have discussed thus far refer to a single job search for worker $i$. In reality, we may observe workers searching for jobs multiple times, and each of these searches is informative about the latent worker skills and job tasks that define worker types $\iota$ and markets $\gamma$. We incorporate repeated searches by assuming that workers periodically receive exogenous separation shocks which arrive following a Poisson process. Upon receiving a separation shock, the worker draws a new $\varepsilon^g_{ij}$ shock and repeats the job choice process described above. Assuming that *Poisson*-distributed exogenous separations happen at a rate $d^g_i$ for the individual worker $i$, then the expected number of times she will match with job $j$ throughout our sample period is given by

$$d^g_i \cdot P(j = j^* | j \in \gamma, i \in \iota, g) = \Omega^g_{\iota\gamma} d^g_i d^g_j. \tag{2.3.4}$$

Equation 2.3.4 forms the basis of our algorithm for clustering workers into worker types and jobs into markets, but before proceeding we must define some notation. Let $N_W$ and $N_J$ denote the number of workers and jobs, respectively, in our data. Define $A_{ij}$ as the number of times that worker $i$ is observed to match with job $j$. Further, define $\boldsymbol{A}$ as the matrix

with typical element $A_{ij}$. $\boldsymbol{A}$ is a $N_W \times N_J$ matrix and represents the full set of worker–job matches observed in our data. As discussed previously, each individual worker belongs to a latent worker type denoted by $\iota$ and each job belongs to a latent market denoted by $\gamma$. The list of all latent worker type and market assignments is stored in the $(N_W + N_J) \times 1$ vector denoted by $\boldsymbol{b}$, known as the *node membership* vector. We define $\boldsymbol{g}$ as the $N_W \times 1$ vector containing each worker's demographic group affiliation. The matrix of worker–job matches $\boldsymbol{A}$ and workers' demographic groups $\boldsymbol{g}$ are the data we use to cluster workers and jobs, while the node membership vector $\boldsymbol{b}$ is the latent object identified by the maximum likelihood procedure we discuss below.

Following equation (2.3.4), the expected number of matches between a worker–job pair, $A_{ij}$, can be written as[15]

$$E[A_{ij}|\boldsymbol{b}, g] = \Omega^g_{\iota\gamma} d^g_i d^g_j. \tag{2.3.5}$$

We prove in Appendix A.7 that our assumption of Poisson-distributed exogenous separation shocks implies that $A_{ij}$ follows a Poisson distribution:

$$A_{ij}|\boldsymbol{b}, g \sim Poisson(\Omega^g_{\iota\gamma} d^g_i d^g_j) \tag{2.3.6}$$

Finally, we incorporate equation (2.3.6) above to fully characterize the likelihood of our data as a function of the unknown parameters, by applying Bayes rule:

$$P(A_{ij}, g|\boldsymbol{b}) = \underbrace{P(A_{ij}|\boldsymbol{b}, g)}_{Poisson(\Omega^g_{\iota\gamma} d^g_i d^g_j)} \underbrace{P(g|\boldsymbol{b})}_{\alpha^g_{\iota\gamma}}, \tag{2.3.7}$$

where $\alpha^g_{\iota\gamma} \equiv P(g|\boldsymbol{b})$ is the fraction of type $\iota$ workers employed in market $\gamma$ jobs who belong to demographic group $g$. Equation 2.3.7 corresponds to a commonly-used method from network theory known as the bipartite degree-corrected stochastic block model with edge weights (SBM). The SBM clusters *nodes* in a network (workers and jobs) into groups (worker types and markets) based on patterns of connections between nodes.[16]. The main parameter

[15]It is worth mentioning that: (i) the information $i \in \iota, j \in \gamma$ is contained in $\boldsymbol{b}$; and (ii) $A_{ij}$ is the number of matches between worker $i$ and job $j$, which makes the event that $j = j^*|i$ equivalent to the event that $A_{ij} = 1$. These two facts allow us to use more succinct notation that directly links theoretical objects in our model to data: $P(j = j^*|j \in \gamma, i \in \iota, g) = P(A_{ij} = 1|\boldsymbol{b}, g)$, which we know the distributional form for. This connects notations from the economic model to the network model, but it still lacks the precise definition of the likelihood of interest, $P(\boldsymbol{A}, \boldsymbol{g}|\boldsymbol{b})$, where $A_{ij}$ can assume values other than just 1.

[16]Larremore et al. (2014) lays out the advantages of using bipartite models over using one-sided network projections to fit SBMs; Karrer and Newman (2011) presents the methodology for degree-correction as it enhances significantly the ability of the SBM to fit large scale real world networks; and Peixoto (2018) deal with weighted SBM inference, which is how I accommodate discrimination influencing matches within the

of interest is the set of assignments of workers to worker types and jobs to markets contained in $\boldsymbol{b}$, while all of the other parameters are nuisance parameters that can be straightforwardly determined after $\boldsymbol{b}$ is defined (Karrer and Newman, 2011). The next step is to maximize the likelihood defined in equation 2.3.7, which we address in the next subsection.

#### 2.3.2.2 A Bayesian approach to recovering worker types and markets

In order to make the estimation of worker types and markets feasible, together with using a principled method for choosing the number of clusters, we employ Bayesian methods from the network literature (Peixoto, 2017). We can rewrite equation (2.3.7) as

$$\begin{aligned} P(\boldsymbol{b}|A_{ij}, g) &\propto P(A_{ij}, g|\boldsymbol{b})P(\boldsymbol{b}) \\ &= \underbrace{P(A_{ij}|\boldsymbol{b}, g)}_{Poisson(\Omega^g_{\iota\gamma} d^g_i d^g_j)} \underbrace{P(g|\boldsymbol{b})}_{\alpha^g_{\iota\gamma}} \underbrace{P(\boldsymbol{b})}_{\text{Prior}} \\ &= P(g|A_{ij}, \boldsymbol{b})P(A_{ij}|\boldsymbol{b})P(\boldsymbol{b}). \end{aligned} \tag{2.3.8}$$

Maximizing the posterior distribution means assigning individual workers to worker types $\iota$ and jobs to markets $\gamma$. The basic intuition follows from and is described in greater detail in Fogel and Modenesi (2022): workers belong to the same worker type if they have approximately the same vector of match probabilities over jobs, while jobs belong to the same market if they have approximately the same vector of match probabilities over workers. The key difference in this paper is that workers in the same worker type $\iota$ may belong to different demographic groups $g$ and each worker type–demographic group pair may face its own wage and therefore have its own match probability. Equation (2.3.8) allows for this by allowing the match probabilities $P(A_{ij}, g|\boldsymbol{b})$ to depend on the workers' demographic group $g$ in addition to the worker types and markets stored in $\boldsymbol{b}$.

If worker types are defined by having common vectors of match probabilities over jobs, but match probabilities are allowed to vary by demographic group within a worker type, how do we know that type $\iota$ workers in group $A$ belong to the same worker type as type $\iota$ workers in group $B$? The answer is embedded in equation (2.3.8). The $\alpha^g_{\iota\gamma}$ term in equation (2.3.8) adjusts workers' match probabilities so that they are relative to their own gender. Suppose women are significantly underrepresented in construction jobs and overrepresented in nursing jobs, and vice versa for men. Once we incorporate this adjustment, we would assign workers to a construction-intensive worker type if they are disproportionately likely to match with construction jobs, *relative to other workers of their gender*. Once we adjust

SBM.

the raw match probabilities to account for this selection, we obtain identical *adjusted* match probability vectors for this group of men and this group of women, causing us to assign them to the same worker type, $\iota$.

Equation (2.3.8) assumes that we know the number of worker types and markets *a priori*, however this is rarely the case in real world applications. Therefore we must choose the number of worker types and markets, $I$ and $\Gamma$ respectively. We do so using the principle of minimum description length (MDL), an information theoretic approach that is commonly used in the network theory literature. MDL chooses the number of worker types and markets to minimize the total amount of information necessary to describe the data, where the total includes both the complexity of the model conditional on the parameters *and* the complexity of the parameter space itself. MDL will penalize a model that fits the data very well but overfits by using a large number of parameters (corresponding to a large number of worker types and markets), and therefore requires a large amount of information to encode it. MDL effectively adds a penalty term in our objective function, such that our algorithm finds a parsimonious model. See Fogel and Modenesi (2022) for greater detail.

Equation (2.3.8) defines a combinatorial optimization problem. If we had infinite computing resources, we would test all possible assignments of workers to worker types and jobs to markets and choose the one that maximizes the likelihood in equation (2.3.8), however this is not computationally feasible for large networks like ours. Therefore, we use a Markov chain Monte Carlo (MCMC) approach in which we modify the assignment of each worker to a worker type and each job to a market in a random fashion and accept or reject each modification with a probability given as a function of the change in the likelihood. We repeat the procedure for multiple different starting values to reduce the chances of finding local maxima. We implement the procedure using a Python package called graph-tool. (https://graph-tool.skewed.de/. See Peixoto (2014a) for details.) Now that we have dealt with the issue of important worker and job characteristics being unobserved, we turn our attention to estimating counterfactuals for wage gap decompositions.

## 2.4 Wage gap decomposition

This section lays out the estimation strategies we use to decompose the gender wage gap, while circumventing some of the issues associated with conventional decomposition methods. We decompose the gender wage gap into the quantities listed in equation (2.2.2): the composition component $E[Y_1(x_{ij})|G_i = 1] - E[Y_1(x_{ij})|G_i = 0]$ and the structural component $E[Y_1(x_{ij}) - Y_0(x_{ij})|G_i = 0]$. The quantity $E[Y_g(x_{ij})|G_i = g] = E[Y_{ij}|G_i = g]$, $g \in \{0, 1\}$ can be consistently and straightforwardly estimated since it is directly observable. The challenge

is estimating the counterfactual wage function $E[Y_1(x_{ij})|G_i = 0]$, given that the potential outcome $Y_1(x_{ij})$ is not observed for female workers. Estimating $E[Y_1(x_{ij})|G_i = 0]$ requires us to use data on male workers to estimate a relationship between observable characteristics $x_{ij}$ and male earnings $Y_1$ and then extrapolate this relationship to female workers.

In this paper, we consider two approaches to estimating counterfactual wage functions. The first is the commonly-used Oaxaca-Blinder decomposition, which we henceforth refer to as OB (Oaxaca, 1973; Blinder, 1973). For the OB decomposition, we estimate two linear regressions — one for the set of male workers and another for the set of female workers — to estimate the functionals $Y_1(\cdot)$ and $Y_0(\cdot)$, respectively, as denoted in equation (2.4.1). Values for $E[Y_g(x_{ij})|G_i = g]$ are obtained by averaging out the fitted values of the respective linear regressions. Estimates for the counterfactual $E[Y_1(x_{ij})|G_i = 0]$ are obtained by using the coefficients from the linear regression fitted for males, $\hat{\beta}_{G=1}$, and multiplying them by the average female covariates, $\bar{x}_{G=0}$, as defined in equation (2.4.1). This is equivalent to producing fitted values for the males' regression, while inputting females' covariates.

$$\begin{aligned}
&\text{OB regressions:} && Y_g(x_i j) = x_{ij}^T \beta_{G=g} + \epsilon_{gij}, \qquad g \in \{0,1\} \\
&\text{OB counterfactual estimate:} && E[Y_1(\widehat{x_{ij}})|G_i = 0] := \bar{x}_{G=0}^T \hat{\beta}_{G=1}, \quad \bar{x}_{G=0} := \sum_{i|G_i=0} \frac{x_{ij}}{n}
\end{aligned} \tag{2.4.1}$$

Our preferred decomposition strategy relies on matching male and female workers with similar observable characteristics and using matched workers of different genders as counterfactuals for each other. This approach was initially proposed by Nopo (2008) and was further extended by Modenesi (2022). Not only does this approach avoid the strong functional form assumptions made by OB, it includes a framework for handling a lack of common support. In this paper, we choose to use the original estimation strategy laid out by Nopo (2008), given its tractability especially for a high-dimensional set of covariates like ours, and we refer to it as the matching decomposition henceforth.

The matching decomposition has two main components: (i) matching observations and (ii) relaxing the overlapping supports assumption. First, counterfactual female earnings $Y_1(x_{ij})|G_i = 0$ — what female workers would have earned if their gender were changed to male but nothing else about them changed — are obtained by *exact matching* each female to one or more male workers with similar observable characteristics and then taking a sample average of the matched males[17]. This method for building counterfactuals is non-parametric,

[17]In this paper we coarsened a few variables such as years of education and age, and we use the coarsened version of these variables instead to perform the exact matching. This serves the purpose of matching more

assuming no functional form for $Y_1(\cdot)$, it exerts no extrapolations out of the support of $x$ and it avoids using data from all workers to build counterfactuals for a specific worker. The matching decomposition handles the lack of common support issue by allowing unmatched workers, i.e. outside of the common support of $x$, to contribute to the overall observed gap. In the matching decomposition, we add two terms, $\Delta_M$ and $\Delta_F$, to the expression for the overall wage gap $\Delta$ in equation (2.2.1) which captures the contributions of unmatched male and female workers, respectively. The resulting expression is

$$\Delta = E[Y_{ij}|G_i = 1] - E[Y_{ij}|G_i = 0] =: \Delta_X + \Delta_0 + \Delta_M + \Delta_F, \tag{2.4.2}$$

where

$$\begin{aligned}
\Delta_X &:= E\left[Y_{ij}|Matched, G_i = 1\right] - E\left[Y_1(x_{ij})|Matched, G_i = 0\right] \\
\Delta_0 &:= E\left[Y_{ij}|Matched, G_i = 1\right] - E\left[Y_1(x_{ij})|Matched, G_i = 0\right] \\
\Delta_M &:= \left\{E\left[Y_{ij}|Unmatched, G_i = 1\right] - E\left[Y_{ij}|Matched, G_i = 1\right]\right\} P\left(Unmatched|G_i = 1\right) \\
\Delta_F &:= \left\{E\left[Y_{ij}|Matched, G_i = 0\right] - E\left[Y_{ij}|Unmatched, G_i = 0\right]\right\} P\left(Unmatched|G_i = 0\right)
\end{aligned}$$

Notice that if all observations are matched the $\Delta_M$ and $\Delta_F$ terms vanish and this method collapses back to the original decomposition we have in equation (2.2.2). The terms $\Delta_X$ and $\Delta_0$ still have the same interpretation as discussed in Section 2.2 — composition and structure, respectively — but now only similar workers of one gender are used to build counterfactuals for the other gender, using an agnostic functional form for the counterfactual function. The extra terms $\Delta_M$ and $\Delta_F$ measure the contribution of unmatched male and female workers to the overall observed gender gap. Each of them measures the difference between matched and unmatched workers of a given gender, weighted by the proportion of unmatched workers within that gender[18]. For example, if unmatched male workers have an average log wage that is 0.2 higher than the average log wage for matched male workers and 10% of male workers are unmatched, then $\Delta_M = 0.2 \times 0.1 = 0.02$.

To understand how the matching decomposition handles a lack of common support, consider male workers employed as professional football players. These workers will not be matched to female workers and therefore would be omitted from the analysis if we simply restricted to the region of common support. However, the male workers do contribute

individuals, giving more statistical power to the method, since workers with just e.g. 1 year difference in age, *ceteris paribus*, are roughly the same in terms of productivity.

[18]Precise definitions of each of the terms in the NP decomposition can be found in the appendix section B.1

meaningfully to the overall gender wage gap because they earn significantly more than the average female worker. The matching decomposition would handle this by including these workers in the $\Delta_M$ term. Intuitively, it would say that some of the gender wage gap can be decomposed within the region of common support, while some of it is explained by male workers outside the region of common support earning more than male workers within the region of common support, and similarly for female workers.

The matching decomposition addresses several limitations inherent in the OB decomposition. First, it relaxes the assumption that $Y_1(x_{ij})$ is linear in $x_{ij}$. Although linear regressions allow for flexible transformations of its covariates, the functional form is still a somewhat arbitrary researcher choice. Second, it allows for heterogeneous returns to covariates across coworkers. Third, the version of the matching estimator we use relaxes the *overlapping supports* assumption, also referred to as the *common supports* assumption. This assumption imposes that the support of $x$ for one of the genders has to fully overlap with the support of $x$ for the other gender, and is imposed by almost all decomposition methods in economics (Fortin et al., 2011). The overlapping supports assumption is imposed to ensure that the counterfactual function $Y_1(x)$ estimated using male data, $x_{G_i=1}$, is only used to predict counterfactual earnings for females whose values of $x$ lie within the male support of $x$. When this condition is not satisfied in the data, observations that are outside of the common support are typically trimmed or given virtually zero weight in the estimation process, potentially eliminating significant numbers of workers from the analysis and making the analysis representative of only a subset of the population (Modenesi, 2022). This is particularly salient when $x$ lies in a high-dimensional space, as is the case in our application with high-dimensional worker types and markets.

Our preferred specifications in this paper use the matching decomposition in conjunction with the latent skills and tasks clusters revealed by our network methodology developed in Section 2.3. Since we define labor market gender discrimination as workers with similar skills performing similar tasks with similar productivity but being paid differently based on gender, our worker type–market clusters serve as natural cells within which workers are considered as equivalent in terms of productivity. With the matching decomposition we are able to ensure that only similar workers are used when estimating counterfactual earnings, mitigating counterfactual biases, and also avoid dropping unmatched workers from the estimation procedure as mentioned above. Although the original matching decomposition is not considered to be a "detailed decomposition" by the literature of decompositions in economics, in combination with our network clusters, it is possible to compute an economically principled distribution of the gender gap (and its components) for a vast amount of cells of workers in the labor market, mapping how discrimination is spread in different

parts of the market.

## 2.5 Data

### 2.5.1 Administrative Brazilian data

We use the Brazilian linked employer-employee data set RAIS. The data contain detailed information on all employment contracts in the Brazilian formal sector, going back to the 1980s. The sample we work with includes all workers between the ages of 25 and 55 employed in the formal sector in the Rio de Janeiro metro area at least once between 2009 and 2018. These workers are defined as matching with the unemployment (or informal sector) in years we do not observe them. We also exclude the public sector because institutional barriers make flows between the Brazilian public and private sectors rare, as well as the military. Finally, we exclude the small number of jobs that do not pay workers on a monthly basis.

Our wage variable is the real hourly log wage in December, defined as total December earnings divided by hours worked. We deflate wages using the national inflation index. We exclude workers who were not employed for the entire month of December because we do not have accurate hours worked information for such workers. We define a job as an occupation-establishment pair. This implicitly assumes that all workers employed in the same occupation at the same establishment are performing approximately the same tasks.

Our data contain 4,578,210 unique workers, 289,836 unique jobs, and 7,940,483 unique worker–job matches. The average worker matches with 1.73 jobs and the average job matches with 27.4 workers. 42% of workers match with more than one job during our sample. Figure 2.1 presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions.

Our network-based classification algorithm identifies 187 worker types ($\iota$) and 341 markets ($\gamma$). Figure 2.2 presents histograms of the number of workers per worker type and jobs per market. The average worker belongs to a worker type with 20,896 workers and the median worker belongs to a worker type with 14,211 workers. The average job belongs to a market with 1,156 jobs and the median job belongs to a market with 1,127 jobs.

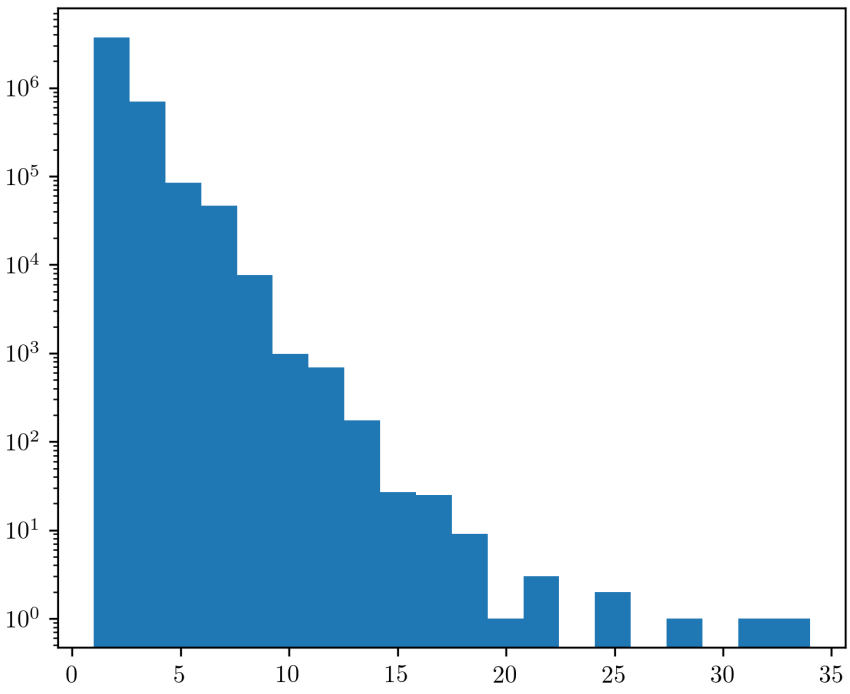
## 2.6 Results

### 2.6.1 Aggregate wage gap decomposition

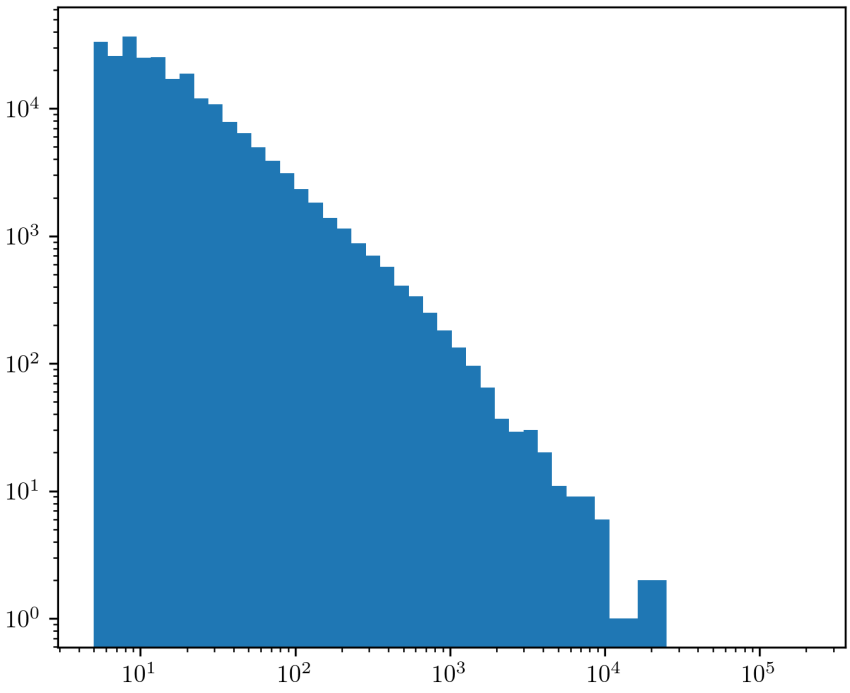
Table 2.1 presents the results of performing gender wage decompositions using each of our two methods: OB and matching. For each method, we have three specifications.

Figure 2.1: Distributions of Number of Matches Per Worker and Job

(a) Workers



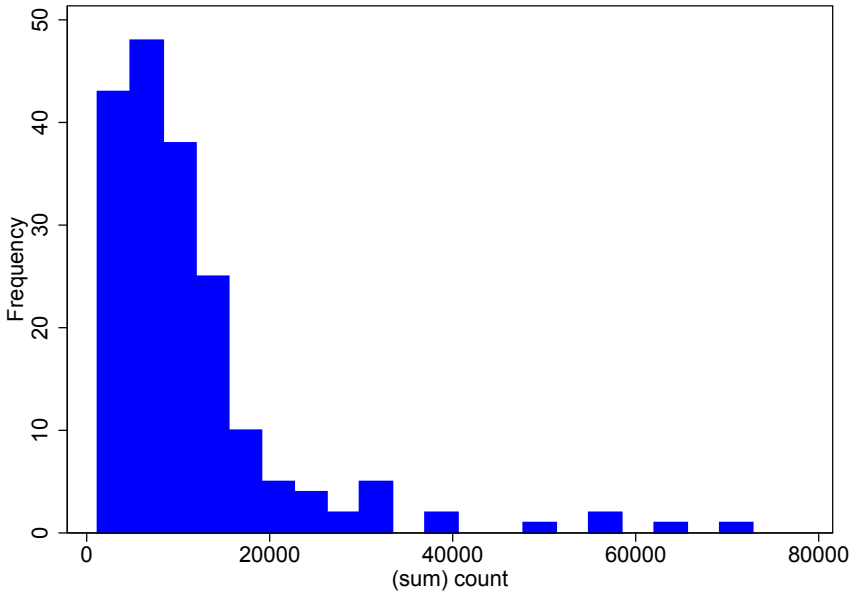

(b) Jobs



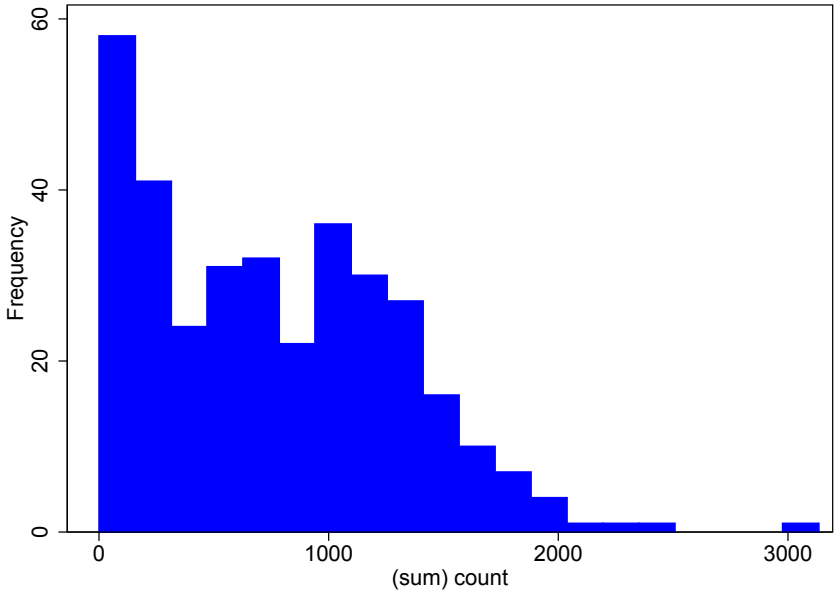

*Notes:* Figure presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions. Vertical axes presented in log scale. Horizontal axis of bottom panel also presented in log scale. Number of matches per worker and job computed from the network of worker–job matches described in Section 2.5.

Figure 2.2: Worker Type ($\iota$) and Market ($\gamma$) Size Distributions

(a) Number of Workers Per Worker Type ($\iota$)




(b) Number of Jobs Per Market ($\gamma$)




*Notes:* Figure presents histograms of the number of workers per worker type $\iota$ and jobs per market ($\gamma$). The units of analysis are worker types in the upper panel and markets in the lower panel. Computed using assignments of workers to worker types and jobs to markets as described in Section 1.3.

The first, presented in columns (1) and (4), estimates counterfactual earnings distributions using a standard set of observable characteristics: experience, education, race, industry and union status. The second, presented in columns (2) and (5), estimates counterfactual earnings distributions using the worker types and markets identified by the SBM. The third specification, presented in columns (3) and (6) uses both standard observable characteristics and worker types and markets. The first row of each column presents the overall wage gap: the average male worker earns 16.7 percent more than the average female worker in our sample. The second row presents the wage gap that would exist if male and female workers with the same productivity were paid equivalently but the observed differences between the distributions of male and female productivity — as proxied by observable characteristics and/or worker types and markets — remained, the *composition* component. The third row presents the wage gap that would exist if male and female workers had identical productivity distributions, but the observed earnings differences conditional on productivity remained, the *structure* component. The fourth and fifth rows present the wage gap explained by male and female workers outside the region of common support, respectively. For the OB method the composition and structure components add up to the overall wage gap; for the matching method the overall wage gap equals the sum of the composition and structure components and the components due to a lack of common support.

The qualitative stories told by both the OB method and the matching method are very similar. When we define counterfactual earnings using observable characteristics (columns 1 and 4), we find that if male and female workers with the same productivity were paid similarly, then female workers would significantly outearn male workers: by 12.7% using the OB method and 8.8% using the matching method. By contrast, female workers are paid significantly less than their male counterparts with similar productivity: 29.4% less using the OB method and 25.6% less using the matching method. When we define counterfactuals using worker types and markets instead of observable characteristics (columns 2 and 5) we find that the wage gap would nearly disappear if male and female workers with the same productivity were paid similarly. By contrast, the wage gap that would exist if male and female workers had the same productivity distribution — 17.9% according to OB and 17.8% according to matching — is almost equal to the overall wage gap of 16.7%. In other words, when we compute counterfactuals using worker types and markets we find that differential pay for similar productivity explains roughly the entire gender wage gap. This tells us that the results of gender wage gap decompositions are highly sensitive to the way in which we define counterfactuals. If, as we argue, worker types and markets do a better job of capturing the latent productivity of worker–job matches than do standard observable characteristics, then these results imply that gender wage gaps are almost entirely due to similarly productive

male and female workers being paid differently, not male and female workers having different productivity distributions.

Columns (3) and (6) of Table 2.1 use both observable characteristics and worker types and markets to form counterfactuals for the gender wage gap decompositions. The OB method finds that female workers have covariates that would imply that they would outearn male workers if equally productive workers were paid equivalently, similar to the findings when we included only observable characteristics, not worker and job types, in column (1). By contrast, the matching method finds that male workers' covariates imply 3.4% higher earnings than female workers' covariates and that male workers are paid 18.5% more than similarly productive female workers. Why do we observe a discrepancy between the OB and matching methods once we include observable characteristics and worker types and markets? The answer lies in the final two rows of Table 2.1, which present the fraction of male and female workers, respectively, for whom we are unable to find a counterfactual. Once we try to match workers on such a large set of variables, many workers are unable to be matched, and a significant part of the gender wage gap occurs among such workers. The matching method allows us to take this into account, while the OB method simply makes a linear extrapolation. However, a linear extrapolation outside the region of common support is likely to lead to incorrect inferences. Furthermore, the fact that the matching estimator yields similar results when we use worker types and markets as it does when we use worker types, markets, and other observable characteristics, but not when we use other observables alone, implies that worker types and markets capture significant determinants of productivity, and omitting them leads to incorrect inferences. This highlights the importance of using a sufficiently set of worker characteristics when estimating counterfactuals, and our method for identifying previously unobserved heterogeneity enhances our ability to do so. All of the results presented in this section correspond to the aggregate gender wage gap. In the next section, we consider heterogeneity in wage gaps within different subsets of the labor market.

### 2.6.2 Wage gaps within worker type–market cells

An appealing feature of our worker types and markets is that they allow us to further decompose gender wage gaps and identify heterogeneity in gender wage gaps across the labor market. We do so by computing overall wage gaps, $\Delta$, and then decomposing them following the matching decomposition, *within* each worker type–market cell.

For each worker type–market cell we decompose the overall wage gap (Row 1 of Table 2.1) into its four components: composition, structure, males unmatched, and females unmatched (Rows 2–5 of Table 2.1). Figure 2.3 presents kernel density plots of the resulting distributions

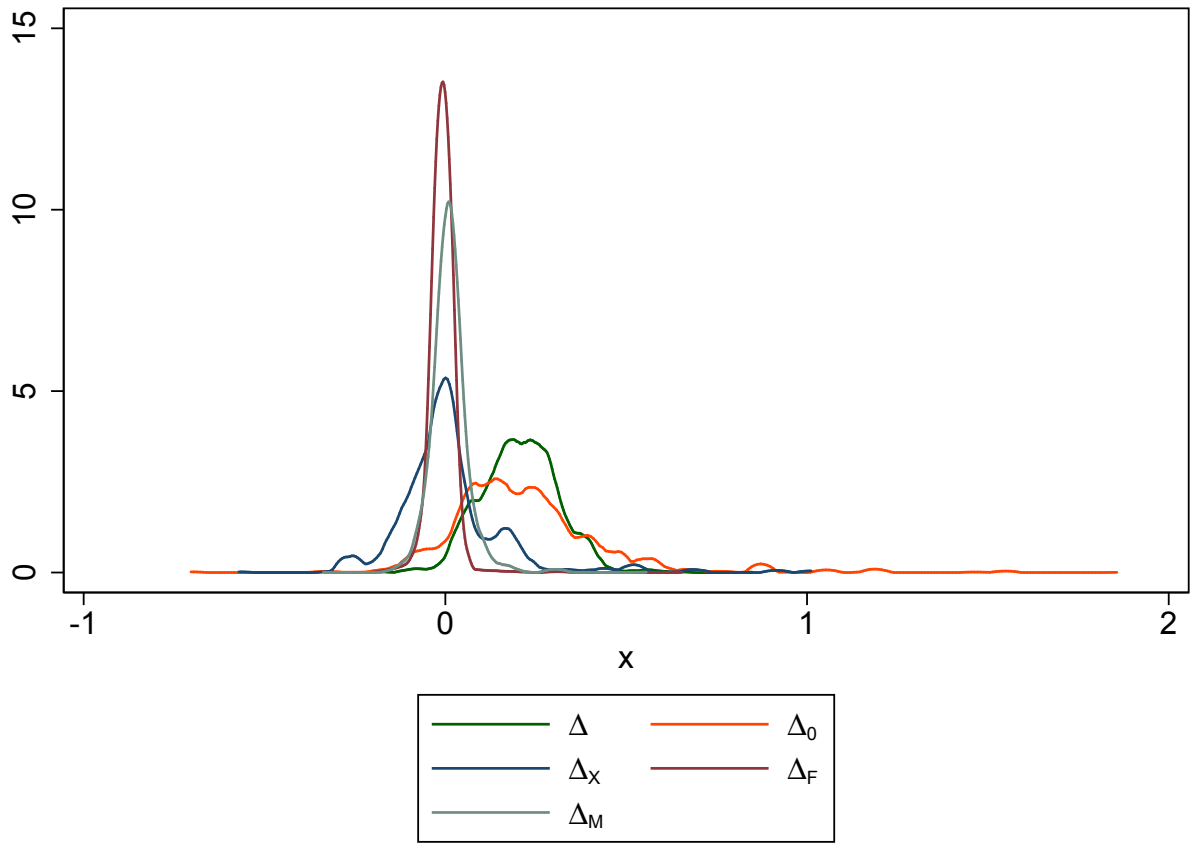
Table 2.1: Gap decomposition using Oaxaca-Blinder vs Matching

| | Oaxaca-Blinder | | | Matching | | |
|---|---|---|---|---|---|---|
| | Observables | $\iota \times \gamma$ | Full model | Observables | $\iota \times \gamma$ | Full model |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Gap | 0.167 | 0.167 | 0.167 | 0.167 | 0.167 | 0.167 |
| Composition | -0.127 | -0.011 | -0.084 | -0.088 | -0.006 | 0.034 |
| Structure | 0.294 | 0.179 | 0.250 | 0.256 | 0.178 | 0.185 |
| Males unmatched | - | - | - | 0.000 | -0.005 | -0.076 |
| Females unmatched | - | - | - | 0.000 | 0.000 | 0.024 |
| % of males matched | - | - | - | 1.00 | 0.98 | 0.57 |
| % of females matched | - | - | - | 1.00 | 0.99 | 0.74 |

*Notes:* All coefficients significant at at least the 1% level.

of overall wage gaps and their four components. Several clear patterns emerge. First, the overall wage gaps $\Delta$ are almost universally positive, meaning that male workers outearning their female counterparts is a widespread phenomenon. Specifically, 91% of workers are in clusters where males outearn females. Second, the distribution of the structural component, $\Delta_0$, is similar to the distribution of the overall wage gap. This suggests that the result from the aggregate decomposition in Section 2.6.1 that almost the entire overall gender wage gap is explained by the structural component holds within worker type–market cells as well. The fact that the structure component roughly coincides with the overall wage gap implies that the other three components — composition, males outside the common support, and females outside the common support — must contribute relatively little to the overall gender wage gap, which is confirmed by the fact that the distributions for these three components are centered close to zero and have low variances. We present the same results quantitatively in Table 2.2. Together, these results tell us that while there is significant variability in gender wage gaps across different worker type–market pairs, the overall qualitative pattern of male workers outearning their female counterparts, and almost all of this gap being explained by differential returns to the same skills rather than different skills, is true in the disaggregated results as well as the aggregated results.

Figure 2.3: Distribution of Components of Overall Wage Gap, Disaggregated




*Notes:* All values measured as differences between male and female log wages. Weighted by number of workers per worker type–market cell. Worker type–market cells with fewer than 1000 workers dropped to remove outliers and improve visual clarity.

Table 2.2: Summary Statistics of Components of Overall Wage Gap, Disaggregated

| | mean | sd | min | max | count |
|---|---|---|---|---|---|
| $\Delta$ | 0.215 | 0.240 | -1.183 | 6.228 | 4791014 |
| $\Delta\_0$ | 0.196 | 0.172 | -2.506 | 9.384 | 4791014 |
| $\Delta\_M$ | 0.016 | 0.134 | -3.577 | 3.448 | 4783255 |
| $\Delta\_F$ | -0.011 | 0.116 | -2.632 | 4.684 | 4724863 |
| $\Delta\_X$ | 0.013 | 0.153 | -1.150 | 2.418 | 4791014 |
| Frac. Male Workers Matched | 0.766 | 0.238 | 0.004 | 1.000 | 4791014 |
| Frac. Female Workers Matched | 0.875 | 0.199 | 0.008 | 1.000 | 4791014 |
| Frac. Workers that Are Male | 0.617 | 0.162 | 0.037 | 0.999 | 4791014 |

## 2.7 Conclusion

In this paper we reconsider the wage gap decomposition literature and make three key contributions. First, we propose a new method for identifying unobserved determinants of workers earnings from the information revealed by detailed data on workerjob matching patterns. The method builds on Fogel and Modenesi (2022) and provides a blueprint for incorporating observable variables into the clustering algorithm, while also relaxing the assumption of perfect competition in labor markets. Second, we non-parametrically estimate counterfactual wage functions for male and female workers and use them to decompose gender wage gaps into a *composition* component in which male and female workers earn different wages because they possess different skills and perform different tasks, and a *structural* component in which male and female workers who possess similar skills and perform similar tasks nonetheless earn different wages. Third, we address the issue of male workers' observables characteristics falling outside the support of female workers' observable characteristics, and vice versa, by augmenting the wage decomposition with components attributable to male and female workers, respectively, outside the region of common support.

We apply these methods to Brazilian administrative data and find that almost the entire gender wage gap is attributable to male and female workers who possess similar skills and perform similar tasks being paid differently. This is true at the aggregate level, and remains true when we perform wage decompositions within each worker type–market cell, indicating that this is a widespread phenomenon, not one driven by large wage differentials in small subsets of the labor market. We find that wage decompositions based on standard observable variables suffer from omitted variable bias, emphasizing the need for detailed worker and job characteristics in the form of worker types and markets. We find that wage decompositions based on linear regressions yield similar findings to those based on matching when a lack of common support is not an issue, however when male and female workers' characteristics do not share a common support the matching estimator with corrections for a lack of common support outperforms alternatives.

While this paper focuses on gender wage gaps, the methods are applicable to other wage gaps, for instance race. Moreover, our strategy for using worker–job matching patterns to control for previously-unobserved, but potentially confounding, covariates may be applied in a wide variety of contexts.

# CHAPTER III

# A Network Theory-Based Attempt to Impute Occupation on the LEHD

## 3.1 Introduction

The U.S. Census Bureau's Longitudinal Employer Household Dynamics (LEHD) is one of the most commonly used administrative data sets by economists. A major limitation of the LEHD is that it contains limited information on individual worker characteristics, such as the worker's occupation, making it difficult for researchers to consider worker skill heterogeneity in their analyses. In this paper I propose and evaluate a method for imputing occupation for workers in the LEHD that uses tools from network theory and machine learning to leverage information contained in the rich network of worker–job matches contained in the LEHD.

The key feature of the data that makes an imputation of occupation possible is that occupation is observed for the subset of workers in the LEHD who also appear in the American Community Survey (ACS). The ACS collects much more detailed worker characteristics than the LEHD, but covers only about 1 percent of the U.S. population every year. The ACS contains the same unique worker identifier as the LEHD, making it straightforward to link workers between the ACS and LEHD. Therefore, my strategy for imputing occupation on the LEHD consists of first modeling the relationship between information contained in the LEHD and occupation as measured in the ACS for the subset of workers who appear in the ACS and then using this relationship to extrapolate occupations to the majority of workers in the LEHD for whom we do not observe occupation.

My approach builds on the insight that the rich network of worker–job matches contained in the LEHD reveals critical information about latent worker and job characteristics. Intuitively, workers tend to match with jobs that involve tasks for which their skills are a good fit. While researchers using the LEHD cannot observe these skills and tasks directly,

workers and jobs can, and they take them into account when forming matches. Therefore, the set of observed matches is informative about worker skill and job task heterogeneity. Every time a worker matches with a job, this indicates that that worker's skills are a good match for that job's tasks. It follows that if two workers match with the same job, those two workers likely have similar skills. Similarly, if a worker moves from job A to job B, I infer that jobs A and B likely require similar tasks. I build upon this intuition to develop a clustering algorithm that identifies groups of workers who tend to match with similar jobs, and thus are revealed to have similar skills, and groups of jobs that tend to hire similar workers and thus are revealed to require similar tasks. I formalize this intuition and cluster workers and jobs into groups, which I denote *worker types* and *markets*, respectively using a method from network theory known as the degree-corrected bipartite stochastic block model (BiSBM). In Chapter I of this dissertation I microfound the BiSBM with a model of workers' labor supply, giving the resulting worker types and markets clear economic interpretability.

A major limitation of my approach is the difficulty of defining a "job." Ideally, a job would represent a set of positions at a firm such that all workers employed in those positions perform approximately the same tasks. Following this definition, it would be reasonable to infer that all workers employed in the same job have similar skills. In Chapter I, we define a job as an establishment–occupation pair, leveraging the fact that the Brazilian RAIS data set allows us to observe each worker's current establishment and occupation. Since I do not observe occupation in the LEHD, it is difficult to distinguish between workers at the same firm performing different tasks. For example, I can't determine whether a worker employed by the University of Michigan is a doctor, lawyer, or cafeteria worker. I attempt to remedy this by defining a job as a firm–earnings bin pair, using a variety of different types of earnings bins. This assumes that workers employed by the same firm who are paid similarly have similar skills. This may work well in settings like a doctor's office where there are clear pay differentials between different types of workers — e.g., doctors, nurses, medical assistants, and receptionists — however it will not work well in settings where workers with distinct skills have similar pay — doctors, lawyers, economists, and engineers at a university. To the extent that I am unable to define jobs in a way that distinguishes between distinct tasks, the worker types and markets I identify will be a noisy indicator of which workers have similar skills and which jobs require similar tasks.

Predicting occupation using only LEHD variables is challenging because the LEHD contains very limited information that can be used to predict workers' occupations: date of birth, place of birth, educational attainment, race, ethnicity, sex, the worker's employer and their quarterly earnings at each employer. Additionally, many observations for some of these variables are imputed, rendering them less useful for predicting occupation. The

network of worker–job matches potentially allows me to circumvent this issue.

I attempt the imputation in two ways: a "naive approach" and a "machine learning approach." In the naive approach I begin by clustering workers into worker types that are revealed to have similar skills based on matching with similar jobs. These worker types are large enough that within each worker type a non-trivial number of workers are included in the ACS and therefore have observable occupations. I use these ACS-linked workers to compute the empirical distribution of occupations for workers in each worker type and then impute occupations for the rest of the workers in that worker type by drawing occupations from the empirical distribution corresponding to that worker type. The machine learning approach similarly estimates a probability distribution of occupations for each worker based on the relationship between occupation and other covariates using the set of workers for whom we observe an occupation in the ACS. However, it estimates the distribution by estimating a random forest model in which a variety of covariates — including worker types, markets, and NAICS industry classification codes — are used as predictors for occupation.

I find that while worker types and markets do have some predictive power for occupations, they have less predictive power than NAICS codes and have very little marginal predictive power when NAICS codes are also included. I also find that the machine learning approach consistently outperforms the naive approach. Specifically, using the machine learning approach with the full set of predictors — including worker types, markets, and NAICS codes — I correctly predict 2-digit occupations 33% of the time. This drops to 22.9% when I exclude NAICS codes from the set of predictors but only to 32.5% when I exclude worker types and markets. By contrast, the naive method predicts 2-digit occupations correctly only 8.8% of the time, while simply drawing 2-digit occupations randomly from the unconditional distribution of 2-digit occupations predicts the correct occupation 6.8% of the time. For 4-digit occupations the analogous prediction accuracies are 14.5%, 13,6%, 9.0%, 1.7%, and 0.1%. Therefore, I conclude that the network of worker–job matches in the LEHD has at best limited predictive power for occupation, however there is enough information about occupations in other variables like NAICS codes that an imputation may still be possible.

While worker types and markets as revealed by the BiSBM are not useful for imputing occupation, they do capture meaningful economic information. I demonstrate that worker types and markets have significant predictive power for earnings that is independent of other standard explanatory variables. Therefore, I conclude that using the BiSBM to cluster workers and jobs based on the network of matches in the LEHD may be a fruitful exercise, however it will need to be applied to applications other than imputing occupations.

## 3.2 Identifying Latent Worker and Job Similarity

The key insight underlying the approach in this chapter is that while researchers using the LEHD are unable to observe many important worker and job characteristics, labor market participants — workers and those at firms who make hiring decisions — do observe worker and job characteristics and base their matching decisions on these characteristics. Therefore, worker–job matching patterns are informative about the set of worker and job characteristics that are unobserved by researchers. Intuitively, every time a worker matches with a job, this is an indication that the worker's skills are probably a good match for the job's tasks. Similarly, if a job hires two different workers we can infer those workers probably have similar skills, while if a worker changes jobs we can infer that the new and old job probably require similar tasks.

I formalize this logic by proposing a model of worker–job matching and use the model to microfound a clustering algorithm that assigns workers and jobs to groups based on the latent skill and task heterogeneity revealed by worker–job matching patterns. A more detailed version of this model is presented in Chapter I of this dissertation. The labor market consists of workers, indexed by $i$, and jobs, indexed by $j$. I assume that workers belong to discrete *worker types*, indexed by $\iota$, such that all workers in the same worker type have the same skills, and all jobs belong to discrete *markets*, indexed by $\gamma$, such that all jobs in the same market consist of the same tasks. I denote the propensity of a type $\iota$ worker to match with a type $\gamma$ job $\mathcal{P}_{\iota\gamma}$ and assume that $\mathcal{P}_{\iota\gamma}$ is constant for all workers in the same worker type $\iota$ and all jobs in the same market $\gamma$. I further assume each job $j$ has a parameter $d_j$ that governs the number of workers it hires. It follows that each time a particular worker $i$ searches for a job, the probability that they end up matched with job $j$ can be written

$$\mathbb{P}_{ij} = \mathcal{P}_{\iota\gamma} d_j \tag{3.2.1}$$

Finally, assume that workers periodically receive Poisson-distributed exogenous job separation shocks that arrive according to an individual-specific frequency parameter $d_i$. Upon receiving such a shock, workers draw a new match according to equation 3.2.1. It follows that the expected number of matches between worker $i$ and job $j$, which we denote $A_{ij}$, follows a Poisson distribution:

$$A_{ij} \sim Poisson\left(d_i d_j \mathcal{P}_{\iota\gamma}\right). \tag{3.2.2}$$

$A_{ij}$ is observable. It is the number of matches between worker $i$ and job $j$ observed in the LEHD. I define the full set of worker–job matches as $\mathbf{A}$, where $A_{ij}$ is the typical element of

**A**.

Since I assume that matches are independent across workers and jobs, I obtain a functional form for the process generating the observed network, encoded in **A** by multiplying the probability in equation (3.2.2) across all workers and jobs:

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{3.2.3}$$

where $\vec{\iota}=\{\iota(i)\}_{i=1}^N$ is the vector assigning each worker to a worker type, $\vec{\gamma}=\{\gamma(j)\}_{j=1}^J$ is the vector assigning each job to a market, $\vec{d}_i=\{d_i\}_{i=1}^N$, $\vec{d}_j=\{d_j\}_{j=1}^J$, and $\mathcal{P}$ is the matrix with typical element $\mathcal{P}_{\iota\gamma}$. From this expression I estimate the worker type and market assignments for all workers and jobs, $\vec{\iota}$ and $\vec{\gamma}$ respectively, using maximum likelihood.

$$\vec{\iota},\vec{\gamma}=\underset{\substack{\{\vec{\iota}=\iota(i)\}_{i=1}^N,\\ \{\vec{\gamma}=\gamma(j)\}_{j=1}^J}}{\arg\max}\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{3.2.4}$$

Equation (3.2.4) assumes that I know the number of worker types and markets *a priori*. Therefore I choose the number of worker types and markets using the principle of minimum description length (MDL), an information theoretic approach that is commonly used in the network theory literature. Equation (3.2.4) corresponds to the degree-corrected bipartite stochastic block model (BiSBM), a workhorse model in the community detection branch of network theory. It defines a combinatorial optimization problem, which I solve using a Markov chain Monte Carlo (MCMC) approach in which I modify the assignment of each worker to a worker type and each job to a market in a random fashion and accept or reject each modification with a probability given as a function of the change in the likelihood. I repeat the procedure for multiple different starting values to reduce the chances of finding local maxima. I implement the procedure using a Python package called graph-tool.[1]

### 3.2.1 Different ways of defining jobs

The clustering algorithm I employ in this dissertation relies on the ability to identify individual workers and jobs. Identifying workers is straightforward because the LEHD contains a unique worker identification number. Identifying jobs is more complicated, both conceptually and practically. Conceptually, I define a job as a set of job titles or positions at a single firm that require workers to perform similar tasks. Ideally, all workers employed by the

[1] https://graph-tool.skewed.de/. See Peixoto (2014a) for details.

same firm to perform similar tasks would be identified as having the same job. In Chapter I, using Brazilian administrative data, we defined a job as an occupation–establishment pair, assuming that workers employed in the same occupation at the same workplace perform similar tasks. This is infeasible in the LEHD because I do not observe occupations for the vast majority of worker–job matches in the LEHD. Therefore, I need some way of distinguishing between different jobs within the same firm.

I define jobs in the LEHD as firm-earnings bin pairs. This implicitly assumes that workers earning similar wages within the same firm are performing similar tasks. This will be a reasonable assumption in firms where different occupations are paid distinct wages. For example, in a small medical office wages may clearly distinguish between doctors, nurses, medical assistants, and administrative staff. This assumption is less plausible in other firms in which distinct jobs are paid similarly. For example, within a large university doctors, lawyers in the general counsel's office, law professors, economics professors, and engineering professors may all be paid similarly. In fact, in some firms there may be more pay heterogeneity within jobs (e.g. within economics professors) than across jobs. To the extent that the data are dominated by firms in which pay does not clearly distinguish between jobs that involve distinct tasks, worker types and markets will do a relatively poor job of identifying workers with similar skills and jobs requiring similar tasks.

I define jobs in four ways based on four different ways of defining within-firm earnings bins. First, I cluster earnings using the Jenks natural breaks classification method. The Jenks method minimizes each cluster's average deviation from the cluster mean, while maximizing each cluster's deviation from the means of the other clusters (Jenks, 1967). The Jenks method can be interpreted as k-means clustering along a single dimension. The Jenks method requires the user to choose the number of bins, so I choose the number of bins as proportional to the natural log of the firm's total employment. Second, I cut each firm's earnings distribution into bins of 10. That is, the 10 lowest-earning workers at the firm have the same job, the 11th-20th lowest-earning workers have the same job, and so on. I call these "quantile bins." Third, I define earnings bins as ventiles (20 quantiles) of the state earnings distribution. Fourth, I ignore earnings altogether and define a job as equivalent to a firm, which I denote "SEIN" in reference to the LEHD's firm ID number, SEIN.

I estimate the BiSBM 4 times — once for each of the different job definitions — and end up with four different sets of worker types, $\{\iota^{Jenks}, \iota^{QuantileBins}, \iota^{Ventiles}, \iota^{SEIN}\}$, and markets, $\{\gamma^{Jenks}, \gamma^{QuantileBins}, \gamma^{Ventiles}, \gamma^{SEIN}\}$.

## 3.3 Data

I begin by creating a worker earnings panel from the LEHD. For computational and confidentiality reasons, I restrict to workers in three states and observations covering 2011–2014. I drop the first and last quarter of employment for job spells that last at least three quarters to avoid quarters in which the worker was not employed for the entire quarter. I keep the first and last quarters for job spells lasting less than three quarters so that I do not lose these workers altogether. I also drop jobs with quarterly mean earnings less than $3770, which is approximately the equivalent of the pay a worker would earn working full time at minimum wage for an entire quarter. I measure earnings as the mean quarterly earnings across all quarters of a job spell that are not dropped for one of the reasons discussed above.

After creating the earnings panel, I merge on occupations from the ACS for the subset of workers in the LEHD who appear in the ACS between 2005 and 2015. For the rare workers who appear in the ACS more than once, I use their most recent occupation. Since the ACS treats occupation as a characteristic of the worker at a point in time, rather than a characteristic of a particular job spell, I am unable to determine which job spell the worker's observed occupation corresponds to with certainty. I therefore treat occupation as a fixed worker characteristic. This allows me to use any ACS occupation observed for that worker, regardless of which year it corresponds to. This has the advantage of allowing me to observe a greater number of ACS occupations, but the disadvantage of adding noise to the relationship between a worker's observed characteristics and job match and their observed occupation.

From the earnings panel I create a data set that represents the network of worker–job matches in the LEHD, which I use for estimating the BiSBM. I define jobs according to the four different definitions explained in Section 3.2.1. I estimate the BiSBM using each of the four job definitions and then merge the estimated worker types and markets back onto the worker earnings panel. Finally, I restrict to one observation per worker, keeping only the longest-tenured job. This gives me the data set I use in the subsequent analysis.

## 3.4 Imputation Attempt

I define occupation in two ways: first using the U.S. Census Bureau's more detailed 4-digit occupation codes ("Occ4") and second using more aggregated 2-digit Standard Occupational Classification (SOC) System codes ("Occ2"). Occ4 provides more detailed information about workers but I also consider Occ2 since it is possible that I will have more success in imputing Occ2.

For both Occ4 and Occ2 I employ two imputation strategies, a "naive approach" and a

"machine learning approach" based on a random forest classifier. I evaluate the success of the imputations using the "accuracy score," which is the ratio of correct predictions to total predictions made.

### 3.4.1 Naive approach

In the naive approach I begin by clustering workers and jobs into worker types and markets, respectively, as described in Section 3.2 and then focus on worker types because I am treating occupation as a worker characteristic. The worker types are large enough that each contains a non-trivial number of workers who appear in the ACS and therefore have observable occupations. I use these ACS-linked workers to compute the empirical distribution of occupations for workers, $\hat{\mathbb{P}}(Occ|\iota)$, for each worker type. I impute occupations for the rest of the workers in worker type $\iota$ by drawing occupations from $\hat{\mathbb{P}}(Occ|\iota)$. Finally, I assess the quality of the imputations by computing the accuracy score for the set of workers for whom I observe an occupation in the ACS.

When I impute Occ2 using the naive approach I obtain an accuracy score of 0.088, meaning that I correctly predict the true occupation 8.8% of the time. A good benchmark for the prediction quality is the accuracy score I would obtain if instead of drawing occupations from the worker type-specific occupation distribution $\hat{\mathbb{P}}(Occ|\iota)$, I instead drew them from the unconditional occupation distribution $\hat{\mathbb{P}}(Occ)$. When I do this I obtain an accuracy score of 0.068. Therefore, the naive imputation strategy based on worker types inferred from the network of worker-job matches using the BiSBM is only marginally better than simply drawing occupations at random. This is consistent with different worker types having similar occupation distributions, and therefore doing a poor job of distinguishing between workers with different skill sets.

I repeat this exercise for 4-digit occupations and obtain accuracy scores of 0.017 and 0.001 when imputing based on $\hat{\mathbb{P}}(Occ|\iota)$ and $\hat{\mathbb{P}}(Occ)$, respectively. The smaller accuracy scores for Occ4 versus Occ2 reflect the fact that there are many more 4-digit occupations (approximately 500 versus 20). When predicting 4-digit occupations, estimated worker types do considerably better than simply using the unconditional Occ4 distribution, however the accuracy is very low, and probably far too low to be useful in empirical applications. Motivated by the poor predictive power of worker types in this naive approach, I turn to a machine learning approach based on a random forest classifier to try to improve predictive power.

### 3.4.2 Machine learning approach

The machine learning approach similarly estimates a probability distribution of occupations for each worker based on the relationship between occupation and other covariates using the set of workers for whom we observe an occupation in the ACS. However, it estimates the distribution by estimating a random forest model in which a variety of covariates are used as predictors for occupation. The full set of predictors consists of worker types and markets inferred from the network of worker–job matches using the BiSBM, employers' 3-digit North American Industrial Classification Codes (NAICS3), workers' quarterly mean earnings, job tenure (in quarters), sex, race, and ethnicity. I consider various specifications in which different subsets of these predictors are used.

The random forest classifier is a commonly-used technique in machine learning that creates a model that predicts the value of an outcome variable based on several input variables, or "features." It builds upon classification trees by training many classification trees and choosing the outcome that is chosen by the most trees. Each classification tree is grown by repeatedly splitting the data based on features in such a way that maximizes the similarity (in terms of the outcome variable) of observations within each of the branches resulting from the split, while minimizing the similarity of observation in different branches. In this way, the classification tree identifies ways of splitting the data that maximize predictive power for the outcome of interest.

The random forest is an aggregation of many classification trees. I draw many different bootstrap samples from my data and fit a classification tree on each sample. Each individual tree in the forest generates a prediction for the class of each observation, and the class with the most votes across all trees in the forest becomes the model's prediction. By averaging over many trees trained on different draws from the same data set, the random forest smooths over the errors present in any individual classification tree, thereby improving predictions and preventing overfitting.

Table 3.1 presents accuracy scores from predicting Occ2 using various subsets of the predictors listed above. All of the worker types and markets are based on the ventiles definition of earnings bins, however results are similar for the other earnings bins (see Tables C.1, C.2, and C.3). The accuracy scores are computed from predictions made using 3-fold cross-validation to prevent overfitting. The first takeaway is that the random forest is able to predict occupations with a much higher accuracy than either the naive method or simply drawing occupations at random. This is reflected by an accuracy of 0.3316 in the random forest using all of the predictors (row 1 of Table 3.1), as compared to 0.088 for the naive method and 0.068 when drawing 2-digit occupations at random. The second takeaway is that worker types and markets have almost no additional predictive power for occupations

relative to NAICS3. We see this by comparing rows 1 and 2 of Table 3.1: when we omit worker types and markets from the set of predictors the accuracy drops from 0.3316 to 0.3215. By contrast, when we use worker types and markets but not NAICS codes as predictors (row 3) the accuracy drops to 0.2285. Using NAICS codes alone (row 12) yields an accuracy of 0.2996. Taken together, these results tell us that while worker types and markets do explain meaningful variation in occupation, NAICS codes are able to explain almost all of this variation, as well as significant variation in occupations that are unexplained by worker types and markets. In other words, worker types and markets tell us almost nothing about occupations that we can not learn from NAICS codes. Therefore, I conclude that worker types and markets are at best minimally useful for imputing occupations in the LEHD. At the same time, these results demonstrate that an occupation imputation based on industry codes may be fruitful.

Table 3.2 presents results analogous to those in Table 3.1, replacing Occ2 with Occ4. The results for Occ4 are qualitatively very similar to those for Occ2. The primary difference is that all of the accuracies for Occ4 are lower, with a maximum of 0.1454 when all predictors are included. The lower accuracies reflect the greater difficulty of predicting more disaggregated occupations. The results for the other three earnings bin definitions are similar (see Tables C.4, C.5, and C.6). Why are worker types and markets less useful than NAICS codes for predicting occupations? I explore this question in the next two subsections.

### 3.4.3 Cramér's V

I attempt to elucidate the findings in Section 3.4 by considering the correlations between some of the relevant variables. I measure correlation between categorical variables using a metric known as "Cramér's V" (Cramér, 1946). Cramér's V is a measure of correlation between two categorical variables that is based on Pearson's chi-squared statistic and ranges from 0 to 1.

Correlations between the different measures of occupation, worker type, and NAICS codes are presented in Table 3.3. The first takeaway, consistent with the finding that NAICS codes have more predictive power for occupations than do worker types, is that NAICS codes are more highly correlated with both Occ2 and Occ4 than are any of the different worker types. Occ2 tends to be more highly correlated with the predictors than Occ4, consistent with the higher accuracy scores for Occ2. Finally, the different worker types are relatively highly correlated with each other, consistent with the fact that accuracy scores varied little depending on which worker type was used. Next, I consider the effectiveness of different predictors in distinguishing between workers with different occupation distributions.

Table 3.1: Accuracy Scores from Predictions of 2-digit Occupation Using Random Forest Classifier

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.3316 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.3251 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.2285 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2180 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2163 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.3057 | Yes | Yes | Yes | No | No | No | No | No |
| 0.1705 | Yes | Yes | No | No | No | No | No | No |
| 0.1738 | No | Yes | No | No | No | No | No | No |
| 0.1543 | Yes | No | No | No | No | No | No | No |
| 0.3104 | No | Yes | Yes | No | No | No | No | No |
| 0.3056 | Yes | No | Yes | No | No | No | No | No |
| 0.2996 | No | No | Yes | No | No | No | No | No |
| 0.2014 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 2-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the ventiles definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

### 3.4.4 Occupation distribution correlations

Worker types and markets are more likely to be useful in imputing occupation if different worker types and markets have very different occupation distributions within them. To understand this, consider the extreme case in which worker types are independent of occupations and therefore carry no information about occupations. Then the distribution of occupations within each worker type would be identical (the empirical distributions may differ somewhat in finite samples due to sampling error) and the correlations between the occupation distributions within different worker types would be close to 1. At the other extreme, if different worker types tend to contain workers with very different occupation distributions, meaning that worker types are informative about occupations, the correlations of occupation distributions across different worker types will be closer to 0. I quantify the amount of information about occupations contained in worker types and markets by computing the occupation distribution within each group and then computing the correlations between these occupation distributions across all pairs of worker types and all pairs of markets. If these correlations tend to be close to 1, then different worker types and markets have similar occupation distributions, meaning that they do a poor job of

Table 3.2: Accuracy Scores from Predictions of 4-digit Occupation Using Random Forest Classifier

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.14540 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.13570 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.08981 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08652 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08477 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.12260 | Yes | Yes | Yes | No | No | No | No | No |
| 0.06160 | Yes | Yes | No | No | No | No | No | No |
| 0.06526 | No | Yes | No | No | No | No | No | No |
| 0.05471 | Yes | No | No | No | No | No | No | No |
| 0.12400 | No | Yes | Yes | No | No | No | No | No |
| 0.12080 | Yes | No | Yes | No | No | No | No | No |
| 0.11500 | No | No | Yes | No | No | No | No | No |
| 0.07589 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 4-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the ventiles definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

discriminating between workers likely to be employed in different occupations.

Figure 3.1 presents kernel density plots of the correlations between the occupation distributions within all pairs of worker types (panel a) and markets (panel b). In each figure there are five lines corresponding to the four different ways of defining jobs — ventiles, Jenks nautral breaks, quantile bins, and SEINs — in addition to 3-digit NAICS codes. If more of the mass of these correlation distributions lies towards the left (closer to correlations of 0) then the relevant worker types and markets are doing a better job of distinguishing between workers in different occupations. For both workers and jobs, we see that the correlations distribution for NAICS codes lies to the left of the distributions for all of the worker types and markets, with the exception of the specification in which jobs are defined as SEINs alone, ignoring earnings bins. This tells us that NAICS codes generally do a better job of distinguishing between workers in different occupations, consistent with the findings in Section 3.4. This is also consistent with the findings in Table 3.3, where we observed that NAICS codes are more highly correlated with occupations than the various worker types. Moreover, Table 3.3 shows that worker types are more highly correlated with NAICS codes when jobs are defined as SEINs than any of the other definitions. This helps explain

Table 3.3: Correlations between Occupations, Worker Types, and NAICS Codes

| | Occ2 | Occ4 | $\iota^{SEIN}$ | $\iota^{Jenks}$ | $\iota^{QuantileBins}$ | $\iota^{Ventiles}$ | NAICS3 |
|---|---|---|---|---|---|---|---|
| Occ2 | 1 | 1 | 0.199 | 0.131 | 0.119 | 0.121 | 0.262 |
| Occ4 | | 1 | 0.089 | 0.083 | 0.174 | 0.086 | 0.215 |
| $\iota^{SEIN}$ | | | 1 | 0.439 | 0.806 | 0.419 | 0.323 |
| $\iota^{Jenks}$ | | | | 1 | 0.788 | 0.354 | 0.172 |
| $\iota^{QuantileBins}$ | | | | | 1 | 0.764 | 0.226 |
| $\iota^{Ventiles}$ | | | | | | 1 | 0.173 |
| NAICS3 | | | | | | | 1 |

*Notes:* Table presents correlations between occupations, worker types (defined based on each of the four different definitions of jobs), and 3-digit NAICS codes. Correlations defined using Cramér's V.

why the correlations distributions for the SEIN-based worker types and markets are more similar to the NAICS codes correlations distribution than are the other worker and job group correlation distributions.

Now that I have established that worker types and markets are poor predictors of occupation, especially relative to 3-digit NAICS codes, I investigate why they are poor predictors and whether or not they carry economic information that is useful in contexts other than imputing occupation.

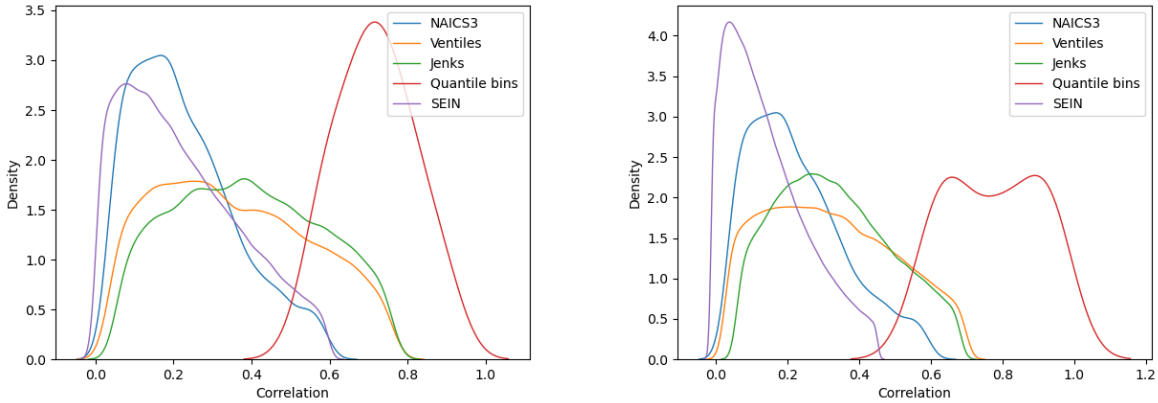
## 3.5 Earnings Regressions

Section 3.4 finds that worker types and markets have insufficient explanatory power for occupation to form the basis of a successful occupation imputation. In this section I consider whether worker types and markets are simply noise, or whether they carry meaningful economic information. I do so by regressing earnings on a variety of covariates, including worker type fixed effects, market fixed effects, and NAICS code fixed effects. I find that worker types and markets do have significant explanatory power for earnings, even conditional on other covariates.

I assess the power of 6 different sets of fixed effects to explain earnings variation. The fixed effects are based on worker types, markets, 3-digit NAICS codes, 4-digit occupation, firms (SEIN), and educational attainment. I estimate a series of regressions, each of which regresses workers' earnings on a different set of fixed effects. There are 6 sets of fixed effects, each of which can be either included or excluded. This yields $2^6 = 64$ different regressions. Each regression also includes the following set of baseline controls: sex, age, age squared, job tenure, and firm employment. I estimate each of these 64 regressions four times — once for each of the four different ways of defining jobs (ventiles, Jenks, quantile bins, and SEINs).

Figure 3.1: Kernel Density Plots of Correlations Between Worker Type and Market Occupation Distributions

(a) Worker Type ($\iota$) (b) Market ($\gamma$)




*Notes:* Panel (a) presents kernel density plots of correlations between occupation distributions within each of the four estimates of worker type as well as 3-digit NAICS codes. Panel (b) presents analogous kernel densities for markets instead of worker types. I use a Gaussian kernel. To preserve confidentiality, the underlying values are rounded to four significant digits and the upper and lower 5% of the distribution are dropped to protect confidentiality.

Similar to the imputation attempts, the results are similar for each of the four different job definitions so for the sake of clarity I focus on the ventiles definition. I present the results in Table 3.4.

Several observations from Table 3.4 stand out. First, all of the sets of fixed effects have at least some predictive power for earnings, relative to the baseline controls. SEIN has the most predictive power, followed by Occ4, market, NAICS3, worker type, and education. Second, while worker type has predictive power for earnings, conditional on including markets it has essentially zero predictive power. That is, any time I add worker type fixed effects to a regression that already includes market fixed effects, the $R^2$ increase is trivial. This is consistent with a story in which because earnings bins do a poor job of distinguishing between jobs in the same firm, our markets approximate clusters of firms, and the worker types approximate clusters of firms plus noise. Markets and 4-digit occupations have similar explanatory power for earnings but they do not explain the same variation, as adding one to a regression that already includes the other significantly increases the $R^2$. Similarly, both worker types and markets explain part of the earnings variation that is unexplained by NAICS codes, and vice versa. Finally, markets add significant explanatory power relative to all other variables, implying that markets identified by the BiSBM do

capture important economic information, however the information they capture is not very predictive of occupations.

Table 3.4: Earnings regressions $R^2$ values

| $R^2$ | $\iota$ | $\gamma$ | NAICS3 | Occ4 | SEIN | Educ |
|---|---|---|---|---|---|---|
| 0.182 | No | No | No | No | No | No |
| 0.213 | No | No | No | No | No | Yes |
| 0.413 | No | No | No | No | Yes | No |
| 0.266 | No | No | Yes | No | No | No |
| 0.324 | No | No | No | Yes | No | No |
| 0.315 | No | Yes | No | No | No | No |
| 0.239 | Yes | No | No | No | No | No |
| 0.432 | No | No | No | No | Yes | Yes |
| 0.297 | No | No | Yes | No | No | Yes |
| 0.413 | No | No | Yes | No | Yes | No |
| 0.329 | No | No | No | Yes | No | Yes |
| 0.492 | No | No | No | Yes | Yes | No |
| 0.379 | No | No | Yes | Yes | No | No |
| 0.332 | No | Yes | No | No | No | Yes |
| 0.489 | No | Yes | No | No | Yes | No |
| 0.356 | No | Yes | Yes | No | No | No |
| 0.409 | No | Yes | No | Yes | No | No |
| 0.263 | Yes | No | No | No | No | Yes |
| 0.437 | Yes | No | No | No | Yes | No |
| 0.302 | Yes | No | Yes | No | No | No |
| 0.359 | Yes | No | No | Yes | No | No |
| 0.317 | Yes | Yes | No | No | No | No |
| 0.432 | No | No | Yes | No | Yes | Yes |
| 0.495 | No | No | No | Yes | Yes | Yes |
| 0.384 | No | No | Yes | Yes | No | Yes |
| 0.492 | No | No | Yes | Yes | Yes | No |
| 0.499 | No | Yes | No | No | Yes | Yes |
| 0.374 | No | Yes | Yes | No | No | Yes |
| 0.489 | No | Yes | Yes | No | Yes | No |
| 0.412 | No | Yes | No | Yes | No | Yes |
| 0.545 | No | Yes | No | Yes | Yes | No |
| 0.438 | No | Yes | Yes | Yes | No | No |
| 0.453 | Yes | No | No | No | Yes | Yes |
| 0.327 | Yes | No | Yes | No | No | Yes |
| 0.437 | Yes | No | Yes | No | Yes | No |
| 0.362 | Yes | No | No | Yes | No | Yes |
| 0.509 | Yes | No | No | Yes | Yes | No |
| 0.401 | Yes | No | Yes | Yes | No | No |
| 0.334 | Yes | Yes | No | No | No | Yes |
| 0.490 | Yes | Yes | No | No | Yes | No |
| 0.358 | Yes | Yes | Yes | No | No | No |
| 0.411 | Yes | Yes | No | Yes | No | No |
| 0.495 | No | No | Yes | Yes | Yes | Yes |
| 0.499 | No | Yes | Yes | No | Yes | Yes |
| 0.547 | No | Yes | No | Yes | Yes | Yes |
| 0.441 | No | Yes | Yes | Yes | No | Yes |
| 0.545 | No | Yes | Yes | Yes | Yes | No |
| 0.453 | Yes | No | Yes | No | Yes | Yes |
| 0.511 | Yes | No | No | Yes | Yes | Yes |
| 0.405 | Yes | No | Yes | Yes | No | Yes |
| 0.509 | Yes | No | Yes | Yes | Yes | No |
| 0.500 | Yes | Yes | No | No | Yes | Yes |
| 0.376 | Yes | Yes | Yes | No | No | Yes |
| 0.490 | Yes | Yes | Yes | No | Yes | No |
| 0.414 | Yes | Yes | No | Yes | No | Yes |
| 0.546 | Yes | Yes | No | Yes | Yes | No |
| 0.439 | Yes | Yes | Yes | Yes | No | No |
| 0.547 | No | Yes | Yes | Yes | Yes | Yes |
| 0.511 | Yes | No | Yes | Yes | Yes | Yes |
| 0.500 | Yes | Yes | Yes | No | Yes | Yes |
| 0.548 | Yes | Yes | No | Yes | Yes | Yes |
| 0.442 | Yes | Yes | Yes | Yes | No | Yes |
| 0.546 | Yes | Yes | Yes | Yes | Yes | No |
| 0.548 | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents $R^2$ values from regressions of workers' earnings on various sets of predictors. All regressions include sex, age, age squared, job tenure, and firm employment. In addition to these baseline controls, each regression includes a different combination of six different fixed effects, each of which can be either included or excluded: worker type ($\iota$), market ($\gamma$), 3-digit NAICS codes, 4-digit occupation, employer ID (SEIN), and education. This yields $2^6 = 64$ different regressions.

## 3.6 Conclusion

This paper attempts to impute occupation on the LEHD by using a method from network theory, the bipartite stochastic block model, to infer latent worker skill and job task information revealed by worker–job matching patterns. It does so by building on the theory developed in Chapter I, which gives the clusters of workers and jobs identified by the BiSBM microfoundation and clear economic interpretability.

I find that the worker types and markets identified by the BiSBM have relatively little explanatory power for occupation, and therefore are unlikely to be the foundation of a successful imputation attempt. An imputation based on 3-digit NAICS industry codes is significantly more accurate than one based on worker types and markets, and adding worker and job types to NAICS codes adds trivial explanatory power. Worker types and markets do have significant explanatory power for workers' earnings, suggesting that they do capture important economic information, however the information they capture is insufficiently predictive of occupation.

# APPENDICES

# APPENDIX A

# Appendix to Chapter 1

## A.1 Adding geography

If we assume the commuting costs are measured in units of our numeraire good, we can add the cost of worker $i$ commuting to job $j$ to the worker's job choice as follows:

$$\gamma_{it} = \underset{\gamma \in \{0,1,\ldots,\Gamma\}}{\arg\max}\ \psi_{\iota\gamma} w_\gamma + \xi_\gamma + CommutingCost_{ij} + \varepsilon_{i\gamma t}$$

Although we have written the commuting cost for a worker $i$ job $j$ pair, we do not observe commuting costs for individual pairs. However, in the market clearing conditions we are integrating over individual workers and jobs of the same type, so really we would only need an integral of commuting costs (basically, average commuting costs).

## A.2 Network theory details

### A.2.1 A primer on networks

"A network is, in its simplest form, a collection of points joined together in pairs by lines" (Newman, 2018). The points are referred to as "nodes", and the lines as "edges." In Figure A.1, the dots represent nodes and the lines represent edges. Networks can represent a wide variety of phenomena. For example, in an air travel network, airports are nodes and flight paths are edges. Similarly, in a social network, people are nodes and edges represent social relationships like friendship. The labor market, as viewed in LEED, can also be represented

as a network. Each node represents an individual worker or job, and each edge represents an employment spell between a worker and a job.

In a network of worker–job connections like ours, edges connect workers to jobs. This means that there can be no edges between two worker nodes or between two job nodes; only between one worker node and one job node. Networks like this, in which nodes belong to one of two categories and all edges connect nodes in different categories, are known as "bipartite" networks. This is reflected in Figure A.1 by the fact that all worker nodes are in blue on the left, all job nodes are in green on the right, and all edges (black lines) connect a worker to a job.

There is one more concept we need to introduce before returning our focus to estimation: the "degree" of a node. The degree of a node is the number of edges connected to that node. In figure A.1, the first (from the top) worker node has a degree of 1 because it is connected to exactly one edge (black line) while the first job node has a degree of 3. We index workers with $i$ and jobs with $j$. We denote the degree of the node representing worker $i$ $d_i$ and the degree of the job representing job $j$ $d_j$. In Figure A.1, $d_{i=1} = 1$ and $d_{j=1} = 3$. As we discuss below, a worker who changes jobs more frequently will have a higher degree, while a job which hires more workers at a given time and/or has higher worker turnover will have a higher degree.

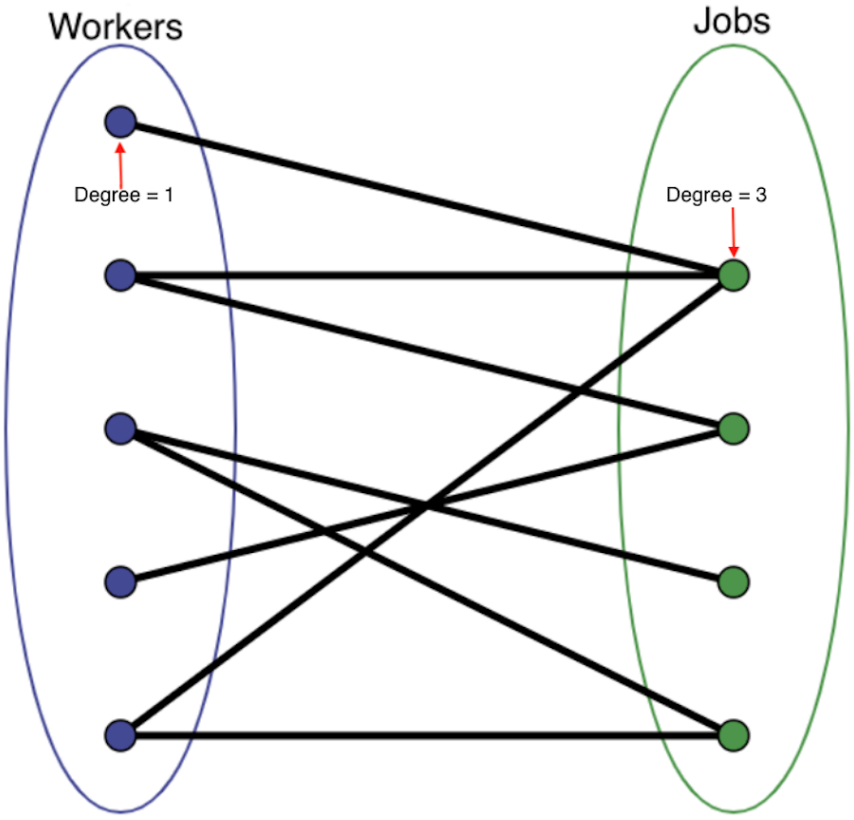
In the next subsection, we show how our model generates a network of worker–job links similar to that in Figure A.1, which can be observed using linked employer-employee data. Then, in the context of our model, we show how to back out latent worker types and markets from this observed network.

### A.2.2 Bipartite Network Details

A network is a collection of nodes (also called "vertices"), connected to each other by edges. A *bipartite* network is a network in which there are two categories of nodes, and all edges connect a node of one category to a node of the other category. In our application, the two categories of nodes are workers and jobs, and all edges connect an individual worker to an individual job. Alternatively, we could have defined a coworker network in which all of the nodes represent individual workers, and an edge connects pairs of workers who are coworkers. The coworker network is not a bipartite network because any node can be connected via an edge to any other node.

One way to represent a network is an adjacency matrix, typically denoted $\mathbf{A}$. The typical element of the adjacency matrix, $A_{ij}$, is the number of edges connecting nodes $i$ and $j$. If there are $n$ nodes in the network, then the adjacency matrix will have dimensions $n \times n$. In equation (A.1) below, we present an adjacency matrix for a bipartite network. Notice

Figure A.1: Simple bipartite network



that there are two large blocks of zeros. This reflects the fact that edges only connect edges of different categories. In our case, edges only connect workers to jobs, not jobs to jobs or workers to workers. Suppose there are $n_J$ jobs andd $n_W$ workers, where $n_J + n_W = n$. Jobs are indexed by $(1, \ldots, n_J)$ and workers by $(n_J + 1, \ldots, n)$.

$$\mathbf{A} = \begin{pmatrix} \overbrace{\begin{matrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{matrix}}^{\text{Jobs}} & \overbrace{\begin{matrix} A_{1,n_J+1} & \ldots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{n_J,n_J+1} & \ldots & A_{n_J,n} \end{matrix}}^{\text{Workers}} \\ \begin{matrix} A_{n_J+1,1} & \ldots & A_{n_J+1,n_J} \\ \vdots & \ddots & \vdots \\ A_{n,1} & \ldots & A_{n,n_J} \end{matrix} & \begin{matrix} 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{matrix} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}}\right\}\text{Jobs} \\ \left.\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}}\right\}\text{Workers} \end{matrix}$$

We can also write the adjacency matrix as

$$\mathbf{A} = \begin{pmatrix} O^{n^J \times n^J} & A^{n^J \times n^W} \\ A^{n^W \times n^J} & O^{n^W \times n^W} \end{pmatrix}$$

where $0^{n \times k}$ is an $n \times k$ matrix of zeros, $A^{n^J \times n^W} = (A^{n^J \times n^W})^T$ and

$$A^{n^J \times n^W} \equiv \begin{pmatrix} A_{1,n_J+1} & \ldots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{n_J,n_J+1} & \ldots & A_{n_J,n} \end{pmatrix}$$

### A.2.3 Stochastic block model details

The *stochastic* in stochastic block model indicates that edges in the network are drawn stochastically from a data generating process (DGP). The *block* refers to the block structure of the DGP. Specifically, the SBM assumes that each node in the network belongs to a group $g \in 1, \ldots, G$. The probability of an edge between two nodes depends solely on group memberships of the two nodes.[1] Therefore, we can write a matrix of edge probabilities that has a block structure:

[1]We have described that standard SBM, as opposed to the degree-corrected version. All of our analysis uses the degree-corrected version, however we ignore that here for simplicity of exposition.

$$
EdgeProbability = \begin{array}{c} \begin{matrix} g(i)=1 & g(i)=1 & g(i)=2 & g(i)=2 \end{matrix} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} \end{array} \begin{matrix} g(i)=1 \\ g(i)=1 \\ g(i)=2 \\ g(i)=2 \end{matrix}
$$

$$
= \begin{pmatrix} p_{g_1,g_1} & p_{g_1,g_1} & p_{g_1,g_2} & p_{g_1,g_2} \\ p_{g_1,g_1} & p_{g_1,g_1} & p_{g_1,g_2} & p_{g_1,g_2} \\ p_{g_2,g_1} & p_{g_2,g_1} & p_{g_2,g_2} & p_{g_2,g_2} \\ p_{g_2,g_1} & p_{g_2,g_1} & p_{g_2,g_2} & p_{g_2,g_2} \end{pmatrix}
$$

In this example, there are four nodes and two groups. Nodes 1 and 2 belong to group 1, as denoted by $g(1) = g(2) = 1$. Similarly, nodes 3 and 4 belong to group 2: $g(3) = g(4) = 2$. Instead of the edge probability matrix above, which can get quite large as the number of nodes grows, we can describe the matrix with two smaller objects: a vector indicating the group assignment of each node and a $G \times G$ matrix of group-specific edge propensities,[2] where $G$ is the number of groups. We denote the vector of group assignments $\vec{g}$ and the matrix of group-specific edge propensities $\mathbf{\Omega}$. then

$$
\vec{g} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}
$$

and

$$
\mathbf{\Omega} = \begin{pmatrix} p_{g_1,g_1} & p_{g_1,g_2} \\ p_{g_2,g_1} & p_{g_2,g_2} \end{pmatrix} \tag{A.1}
$$

Now we describe how to generate a network using the stochastic block model, given parameters. Let $\mathbf{A}$ be the adjacency matrix of a network with $n = 4$ nodes and $\vec{g}$ and $\mathbf{\Omega}$ described above, with $\omega_{rs}$ representing an element of $\mathbf{\Omega}$. We assume that edges are placed between each pair of nodes, $i$ and $j$, following a Poisson distribution with mean equal to the edge probability corresponding to the nodes' respective groups: $\omega_{g_i,g_j}$. Therefore, the

[2] These are not technically probabilities but they can be normalized to be probabilities.

probability of drawing $A_{ij}$ edges between nodes $i$ and $j$ is

$$P(A_{ij}|\omega_{g_ig_j}, g_i, g_j) = \frac{(\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\omega_{g_ig_j}\right).$$

The probability is slightly different for self-edges (edges connecting a node to itself):[3]

$$P(A_{ii}|\omega_{g_ig_i}, g_i) = \frac{(\frac{1}{2}\omega_{g_ig_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}\omega_{g_ig_i}\right).$$

The probability of observing the entire network, represented by $\mathbf{A}$, is the product of the probabilities of each element in the adjacency matrix:

$$P(\mathbf{A}|\mathbf{\Omega}, \vec{g}) = \prod_{i<j} \frac{(\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\omega_{g_ig_j}\right) \times \prod_i \frac{(\frac{1}{2}\omega_{g_ig_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}\omega_{g_ig_i}\right) \tag{A.2}$$

Although equation (A.2) presents the standard SBM, this formulation is rarely used in practice. For empirical applications, researchers typically use an extension called the *degree-corrected* stochastic block model (DCSBM). The difference between the SBM and the DCSBM is that the DCSBM allows the expected degree of each node (the number of edges connected to that node) to vary. This more-closely matches real world data and the DCSBM has been shown to have far superior performance in empirical applications than the SBM (Karrer and Newman, 2011). Let $\vec{d}$ be vector containing the degree of each node, with typical element $d_i$ representing the degree of node $i$. We can write the DCSBM as

$$P(\mathbf{A}|\vec{d}, \mathbf{\Omega}, \vec{g}) = \prod_{i<j} \frac{(d_i d_j \omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_i d_j \omega_{g_ig_j}\right) \times \prod_i \frac{(\frac{1}{2}d_i^2\omega_{g_ig_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}d_i^2\omega_{g_ig_i}\right). \tag{A.3}$$

### A.2.4 Community detection using the stochastic block model

In Section A.2.3 we assumed that we know all of the parameters of the model: $\vec{d}$, $\mathbf{\Omega}$, and $\vec{g}$. However, in actual applications, we typically observe the network $\mathbf{A}$ and the degree distribution $\vec{d}$ and want to recover the group memberships of the nodes $\vec{g}$. (Conditional on knowing $\vec{g}$, we can also compute the empirical edge probabilities matrix $\hat{\mathbf{\Omega}}$.) Therefore, we recover the group memberships of the nodes, $\vec{g}$, by treating equation (A.3) as a maximum likelihood problem and choosing the group memberships in order to maximize the probability

[3]For more details, see section II of Karrer and Newman (2011).

of the observed adjacency matrix $\mathbf{A}$, given the data. We write the likelihood

$$\mathcal{L}(A|\vec{g}) = \prod_{i<j} \frac{(d_i d_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_i d_j \omega_{g_i g_j}\right) \times \prod_i \frac{(\frac{1}{2} d_i^2 \omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2} d_i^2 \omega_{g_i g_i}\right) \quad \text{(A.4)}$$

and our task is to choose

$$\hat{\vec{g}} = \arg\max_{\vec{g}} \mathcal{L}(A|\vec{g})$$

### A.2.5 Bipartite stochastic block model details

The bipartite stochastic block model (BiSBM) is an extension of the SBM (Section A.2.3) applied to bipartite networks (Section A.2.2). The edge probability matrix has the same block structure as in the SBM, however since it is a bipartite network, there are two categories of nodes — in our case workers and jobs — and all edges connect a node from one category (a worker) to a node from the other (job).

Suppose there are two types of workers, indexed by $\iota \in 1, 2$, and two types of jobs, indexed by $\gamma \in 1, 2$. Suppose further that there are 4 individual workers and 4 individual jobs, indexed by $i = 1, \ldots, 4$ and $j = 1, \ldots, 4$, respectively. There are two individual workers and two individual jobs of each type. Denote the probability of an edge between a type $\iota$ worker and a job in market $\gamma$ as $\omega_{\iota\gamma}$. Then we have the following edge probability matrix

| | Jobs | | | | Workers | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $i=1$ | $i=2$ | $i=3$ | $i=4$ | }Worker/Job Index |
| | $\gamma=1$ | $\gamma=1$ | $\gamma=2$ | $\gamma=2$ | $\iota=1$ | $\iota=1$ | $\iota=2$ | $\iota=2$ | }Worker/market |
| $j=1,\gamma=1$ | 0 | 0 | 0 | 0 | $\omega_{11}$ | $\omega_{11}$ | $\omega_{21}$ | $\omega_{21}$ | |
| $j=2,\gamma=1$ | 0 | 0 | 0 | 0 | $\omega_{11}$ | $\omega_{11}$ | $\omega_{21}$ | $\omega_{21}$ | }Jobs |
| $j=3,\gamma=2$ | 0 | 0 | 0 | 0 | $\omega_{12}$ | $\omega_{12}$ | $\omega_{22}$ | $\omega_{22}$ | |
| $j=4,\gamma=2$ | 0 | 0 | 0 | 0 | $\omega_{12}$ | $\omega_{12}$ | $\omega_{22}$ | $\omega_{22}$ | |
| $i=1,\iota=1$ | $\omega_{11}$ | $\omega_{11}$ | $\omega_{12}$ | $\omega_{12}$ | 0 | 0 | 0 | 0 | |
| $i=2,\iota=1$ | $\omega_{11}$ | $\omega_{11}$ | $\omega_{12}$ | $\omega_{12}$ | 0 | 0 | 0 | 0 | }Workers |
| $i=3,\iota=2$ | $\omega_{21}$ | $\omega_{21}$ | $\omega_{22}$ | $\omega_{22}$ | 0 | 0 | 0 | 0 | |
| $i=4,\iota=2$ | $\omega_{21}$ | $\omega_{21}$ | $\omega_{22}$ | $\omega_{22}$ | 0 | 0 | 0 | 0 | |

The primary takeaway from this matrix is that the probability of a connection between a pair of nodes is determined by their group memberships. If worker $i$ belongs to type $\iota$ and

job $j$ belongs to type $\gamma$, then the probability of worker $i$ matching with job $j$ is governed by $\omega_{\iota\gamma}$. The two blocks of zeros in this matrix reflect the fact that the probability of an edge between two workers or two jobs is zero in a bipartite network.

We can write the DGP for the BiSBM as we did above for the standard or degree-corrected SBM. Here we will use the degree-corrected version, since that is what we use for estimation The probability of $A_{ij}$ edges between worker $i$ and job $j$ is given by

$$P(A_{ij}|\omega_{g_i g_j}, g_i, g_j, d_i, d_j) = \frac{(d_i d_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_i d_j \omega_{g_i g_j}\right)$$

From this, we can compute the likelihood of the full observed network, represented by the adjacency matrix $\mathbf{A}$. However, it is inmportant to note that the product below is only over pairs of nodes that *belong to opposite categories*. That is, if $i$ indexes workers and $j$ indexes jobs, we are only taking the product over $i, j$ pairs, not $i, i'$ or $j, j'$ pairs. Again, this is because in a bipartite network, edges can only connect nodes that belong to different categories.

$$P(\mathbf{A}|\vec{d}, \mathbf{\Omega}, \vec{g}) = \prod_{i<j} \frac{(d_i d_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_i d_j \omega_{g_i g_j}\right). \tag{A.5}$$

Notice that this expression lacks the second term found in equation (A.3), which captures self-edges in which an edge runs connects a node to itself. This is because self-edges are impossible in a bipartite network, since self-edges would connect nodes belonging to the same category (e.g. workers to workers).

### A.2.6 Visual representation of linked employer-employee data as a network

Our raw data looks like what is presented in Table A.1, with the exception that we generate the "JobID" column ourselves by concatenating the establishment code ('Estab Code') and occupation code ('Occ Code'). However, we only use the two variables 'WorkerID' and 'JobID' in estimation. Therefore, in Figure A.2, we show the worker and job IDs from the data alongside a network representation of the same data. In the network representation, workers are blue dots on the right, jobs are yellow dots on the left, and black lines represent edges connecting workers to jobs at which they were employed. Finally, in Table A.2, we present an adjacency matrix representation of the same network.

Table A.1: Sample linked-employer-employee data

| WorkerID | Establishment | Occupation | Estab Code | Occ Code | JobID |
|---|---|---|---|---|---|
| 1 | Walmart | Cashier | 1 | 1 | 1_1 |
| 2 | Walmart | Cashier | 1 | 1 | 1_1 |
| 2 | Kroger | Cashier | 2 | 1 | 2_1 |
| 3 | Walmart | Cashier | 1 | 1 | 1_1 |
| 3 | Walmart | Greeter | 1 | 2 | 1_2 |
| 4 | Walmart | Greeter | 1 | 2 | 1_2 |
| 5 | Walmart | Cashier | 1 | 1 | 1_1 |
| 5 | Kroger | Cashier | 2 | 1 | 2_1 |
| 6 | Walmart | Greeter | 1 | 2 | 1_2 |
| 6 | CVS | Manager | 3 | 3 | 3_3 |
| 6 | Chipotle | Manager | 4 | 3 | 4_3 |
| 7 | Chipotle | Manager | 4 | 3 | 4_3 |
| 8 | CVS | Manager | 3 | 3 | 3_3 |
| 8 | Chipotle | Manager | 4 | 3 | 4_3 |
| 9 | Chipotle | Manager | 4 | 3 | 4_3 |
| 9 | Kroger | Asst. Mgr | 2 | 5 | 2_5 |
| 10 | CVS | Manager | 3 | 3 | 3_3 |
| 10 | Chipotle | Manager | 4 | 3 | 4_3 |
| 10 | Chili's | Waiter | 5 | 4 | 5_4 |
| 10 | Kroger | Asst. Mgr | 2 | 5 | 2_5 |

Figure A.2: Representing the data as a network

| WorkerID | JobID |
|---|---|
| 1 | 1_1 |
| 2 | 1_1 |
| 2 | 2_1 |
| 3 | 1_1 |
| 3 | 1_2 |
| 4 | 1_2 |
| 5 | 1_1 |
| 5 | 2_1 |
| 6 | 1_2 |
| 6 | 3_3 |
| 6 | 4_3 |
| 7 | 4_3 |
| 8 | 3_3 |
| 8 | 4_3 |
| 9 | 4_3 |
| 9 | 2_5 |
| 10 | 3_3 |
| 10 | 4_3 |
| 10 | 5_4 |
| 10 | 2_5 |

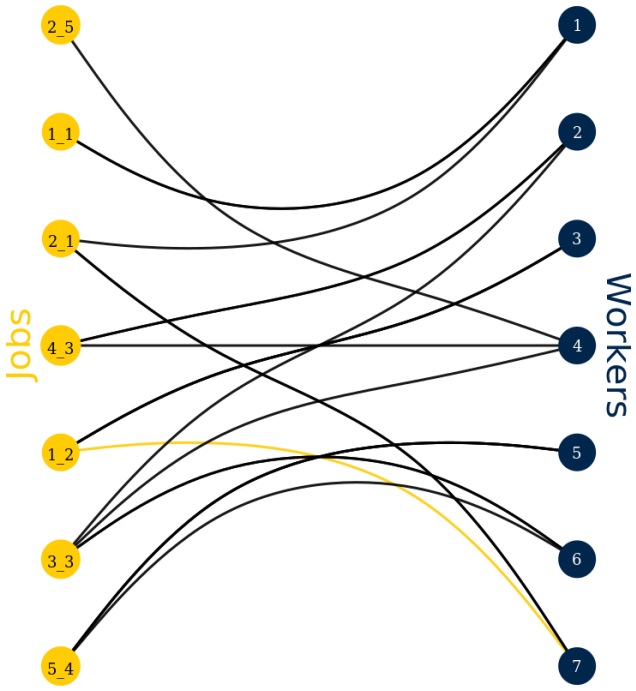


Table A.2: Adjacency matrix: **A**

| **Worker \Job** | **1_1** | **1_2** | **2_1** | **2_5** | **3_3** | **4_3** | **5_4** |
|---|---|---|---|---|---|---|---|
| **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **3** | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **5** | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **6** | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **8** | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| **9** | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| **10** | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

## A.3 Model Solution Appendix

**Firm's problem**

*This section describes a slightly different version of the firm's problem than we presented in the body of the paper. In the body of the paper we had a set of competitive firms in each sector, whereas in what follows here we have a single representative firm in each sector.*

$$\max_{\ell_{\gamma s}} \quad p_s \prod_{\gamma} \ell_{\gamma s}^{\beta_{\gamma s}} - \sum_{\gamma} w_\gamma \ell_{\gamma s} \tag{A.1}$$

There are $S$ optimizations with $\Gamma$ choice variables each, giving us $S \times \Gamma$ FOCs.
FOC:

$$\ell_{\gamma s}^D = \frac{p_s \beta_{\gamma s} \left( \prod_{\gamma'} \ell_{\gamma' s}^{D\ \beta_{\gamma' s}} \right)}{w_\gamma} \tag{A.2}$$

Combining the $\Gamma$ FOCs for a given sector $S$:

$$\ell_{\gamma s}^D = \frac{\beta_{\gamma s}}{\beta_{\gamma' s}} \frac{w_{\gamma'}}{w_\gamma} \ell_{\gamma' s}^D \tag{A.3}$$

Plugging in A.3 for $\ell_{\gamma s}^D$ in equation A.2, we have

$$\ell_{\gamma s}^D = \left[ p_s \left( \frac{\beta_{\gamma s}}{w_\gamma} \right)^{1-\sum_{\gamma'} \beta_{\gamma' s}} \prod_{\gamma'} \left( \frac{\beta_{\gamma' s}}{w_{\gamma'}} \right)^{\beta_{\gamma' s}} \right]^{\frac{1}{1-\sum_{\gamma'} \beta_{\gamma' s}}} = \ell_{\gamma s}^D \left( \vec{p}, \vec{w} \right) \tag{A.4}$$

which represents labor demand for firm $s$, using only FOCs for firm $s$.[4]

Since labor is the only factor of production, we can write firm $s$'s product market supply as

$$y_s^S = y_s^S \left( \{ \ell_{\gamma s}^D (\vec{p}, \vec{w}) \}_{\gamma=1}^{\Gamma} \right) = \prod_{\gamma} \ell_{\gamma s}^{D\ \beta_{\gamma s}} \tag{A.5}$$

**Household's problem**

[4]We could alternatively write this expression as

$$\ell_{\gamma s}^D = \left( \frac{\beta_{\gamma s}}{w_\gamma} \right) \left[ p_s \prod_{\gamma'} \left( \frac{\beta_{\gamma' s}}{w_{\gamma'}} \right)^{\beta_{\gamma' s}} \right]^{\frac{1}{1-\sum_{\gamma'} \beta_{\gamma' s}}}$$

$$\max_{\{y_s^D\}_{s=1}^S} \quad \underbrace{\left(\sum_s a_s^{\frac{1}{\eta}} y_s^{D\,\frac{\eta-1}{\eta}}\right)^{\frac{\eta}{\eta-1}}}_{U(\{y_s^D\}_{s=1}^S)} \quad \text{s.t.} \quad \sum_s p_s y_s \leq Y$$

Lagrangean:

$$\underbrace{\left(\sum_s a_s^{\frac{1}{\eta}} y_s^{D\,\frac{\eta-1}{\eta}}\right)^{\frac{\eta}{\eta-1}}}_{U(\vec{y}^D)} - \lambda\left(\sum_s p_s y_s - Y\right)$$

FOC:

$$\frac{\eta}{\eta-1} U^{\frac{1}{\eta}} \frac{\eta-1}{\eta} a_s^{\frac{1}{\eta}} y_s^{D\,-\frac{1}{\eta}} - \lambda p_s = 0$$

Simplifying,

$$U^{\frac{1}{\eta}} a_s^{\frac{1}{\eta}} y_s^{D\,-\frac{1}{\eta}} - \lambda p_s = 0$$

Rearranging,

$$y_s^D = \frac{U}{\lambda^\eta} \frac{a_s}{p_s^\eta} \tag{A.6}$$

Next, we plug this into the constraint satisfied with equality ($\sum_s p_s y_s^D = Y$):

$$\frac{U}{\lambda^\eta} \sum_s \left(a_s p_s^{1-\eta}\right) = Y$$
$$\Rightarrow \lambda^\eta = \frac{U}{Y} \sum_{s'} \left(a'_s p_s'^{1-\eta}\right)$$

Plugging this into A.6, we have our expression for product demand:

$$y_s^D = \frac{a_s Y}{p_s^\eta \sum_{s'} \left(a_{s'} p_{s'}^{1-\eta}\right)} = y_s^D(\vec{p}, Y) \tag{A.7}$$

**Worker's problem**

$$\max_{\gamma} \quad w_\gamma \psi_{\iota\gamma} + \xi_\gamma + \varepsilon_{i\gamma}, \quad \varepsilon_{i\gamma} \sim T1EV(\theta)$$

Solving the worker's problem gives labor supply:

$$\ell_\gamma^S(\vec{w}) = \sum_\iota m_\iota \left( \frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma} \tag{A.8}$$

**Equilibrium**

Equilibrium wages $\vec{w}_{\Gamma\times 1}$ and prices $\vec{p}_{S\times 1}$ must satisfy three market clearing conditions:

1. Labor market:

$$\sum_s \ell_{\gamma s}^D = \ell_g^S \quad \forall \gamma \in \{1, \ldots, \Gamma\}$$

2. Product market:

$$y_s^D = y_s^S \quad \forall s \in \{1, \ldots, S\}$$

3. Spending = Income = Wages + Profits

$$Y \equiv \sum_s p_s y_s^D = W + \Pi \equiv \sum_s p_s y_s^S$$

where

1. Product demand:

$$y_s^D = \frac{a_s Y}{p_s^\eta \sum_{s'} \left(a_{s'} p_{s'}^{1-\eta}\right)}$$

2. Product supply:

$$y_s^S = \prod_\gamma {\ell_{\gamma s}^D}^{\beta_{\gamma s}}$$

3. Labor supply:

$$\ell_\gamma^S(\vec{w}) = \sum_\iota m_\iota \left( \frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma}$$

4. Labor demand:

$$\ell_{\gamma s}^{D} = \left[ p_s \left( \frac{\beta_{\gamma s}}{w_\gamma} \right)^{1-\sum_{\gamma'} \beta_{\gamma' s}} \prod_{\gamma'} \left( \frac{\beta_{\gamma' s}}{w_{\gamma'}} \right)^{\beta_{\gamma' s}} \right]^{\frac{1}{1-\sum_{\gamma'} \beta_{\gamma' s}}}$$

5. Budget (which can be plugged in for $Y$ in the product demand equation)

$$Y = \sum_s p_s y_{\gamma s}^{S}$$

This is enough for equilibrium, which we find numerically using fixed point iteration. The algorithm proceeds as follows:

1. Choose vectors of start values for wages $\vec{w}$ and prices $\vec{p}$
2. Compute labor supply $\ell_\gamma^S(\vec{w})$ given wages $\vec{w}$ following equation A.8
3. Compute labor demand $\ell_{\gamma s}^{D}(\vec{p}, \vec{w})$ given these start values following equation A.4
4. Compute the product supply $y_s^S\left(\{\ell_{\gamma s}^{D}(\vec{p}, \vec{w})\}_{\gamma=1}^{\Gamma}\right)$ implied by the labor demand choice in the previous step following equation A.5
5. Compute household income $Y = \sum_s p_s y_{\gamma s}^{S}$ implied by product supply in the previous step
6. Compute product demand $y_s^D(\vec{p}, Y)$ following equation A.7
7. Update prices using the update rule $p_s^{t+1} = p_s^t \left(\frac{y_s^D}{y_s^S}\right)^\rho$, where $\rho$ is a dampening factor that controls the size of the update and $t$ indexes iterations. Intuitively, we increase prices if demand exceeds supply, and decrease them if supply exceeds demand. The size of the update depends on the size of the mismatch between supply and demand.
8. Update wages using the update rule $w_\gamma^{t+1} = w_\gamma^t \left(\frac{\ell_\gamma^D}{\ell_\gamma^S}\right)^\rho$
9. Repeat steps 2-8 until convergence

## A.4 Choosing number of worker types and markets

Equation 3.2.3 defined the probability of observing our network of worker–job matches, denoted by the adjacency matrix $\mathbf{A}$:

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right). \tag{A.1}$$

As Peixoto (2017) shows, we can think of this in Bayesian terms and write the full joint distribution of the data, $\mathbf{A}$, and the parameters, $\vec{\iota}$, $\vec{\gamma}$, $\vec{d}_i$, and $\vec{d}_j$ as

$$P\left(\mathbf{A},\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)P\left(\vec{d}_i,\vec{d}_j\middle|\vec{\iota},\vec{\gamma},\mathcal{P}\right)P\left(\mathcal{P}\middle|\vec{\iota},\vec{\gamma}\right)P\left(\vec{\iota},\vec{\gamma}\right) \tag{A.2}$$

where $P\left(\vec{d}_i,\vec{d}_j\middle|\vec{\iota},\vec{\gamma},\mathcal{P}\right)$, $P\left(\mathcal{P}\middle|\vec{\iota},\vec{\gamma}\right)$, and $P\left(\vec{\iota},\vec{\gamma}\right)$ are prior probabilities.

It turns out that this Bayesian formulation has an equivalent information-theoretic interpretation. We can rewrite the joint probability defined in equation (A.2) as

$$P\left(\mathbf{A},\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=2^{-\Sigma}$$

where

$$\Sigma=-\log_2 P\left(\mathbf{A},\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=\mathcal{S}+\mathcal{L}$$

is called the description length of the data and represents the number of bits necessary to encode the data.

$$\mathcal{S}=-\log_2 P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)$$

represents the number of bits necessary to encode the model, conditional on knowing the model parameters, and

$$\mathcal{L}=-\log_2 P\left(\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)$$

is the number of bits necessary to encode the model parameters. $\mathcal{S}$ will be small if the model fits the data well, and $\mathcal{L}$ will be small if the complexity of the model (in our case, the number of worker types and markets) is small. This implicitly defines a trade-off. As we

add more worker types and markets, the model fits the data better, reducing $\mathcal{S}$; however, we are increasing the complexity of the model and thereby increasing $\mathcal{L}$. MDL resolves this trade-off by minimizing $\mathcal{S} + \mathcal{L}$.

We choose the assignment of workers to worker types and jobs to markets that maximizes the posterior of the distribution, equation (A.2). This is equivalent to choosing the set of parameters that yields the smallest description length, and therefore compresses the data the most. Intuitively, we can think of $\mathcal{L}$ as a penalty term that increases with the number of parameters, and thereby prevents overly complex models. If the number of worker types and markets becomes large, $\mathcal{S}$ will increase, indicating a better model fit, but the penalty term $\mathcal{L}$ will increase as well. The chosen model will therefore be the one that maximizes the quality of the model fit relative to the cost imposed by the penalty term.

For more detail, see Peixoto (2014b) and Gerlach et al. (2018).

## A.5 Identification of Labor Supply Parameters

Taking the first order conditions of equation 1.4.3 with respect to each of the parameters provides intuition for how the parameters are identified.

### A.5.1 $\nu$

$$\ell_\nu = 0 \Rightarrow \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\left[\sum_{\gamma'} \mathbb{P}(\gamma'|\Theta)(\phi_{\iota_i\gamma'} + \xi_{\gamma'}) - (\phi_{\iota\gamma_{it}} + \xi_{\gamma_{it}})\right] = 0$$

Intuitively, $\nu$ will be larger if more workers' actual market choices deviate from the choice those workers would have made in the absence of the preference shock $\varepsilon$. The first term in the bracket, $\sum_{\gamma'} \mathbb{P}(\gamma'|\Theta)(\phi_{\iota_i\gamma'} + \xi_{\gamma'})$ is the expected systematic (excluding the idiosyncratic component, $\varepsilon$) utility of the optimal market choice for worker $i$ and, and the second term, $\phi_{\iota\gamma_{it}} + \xi_{\gamma_{it}}$ is the systematic utility for worker $i$ in the market they actually chose in period $t$. Intuitively, if this difference is large, it must be because some workers received large idiosyncratic preference shocks, $\varepsilon_{i\gamma t}$, which caused them to accept otherwise suboptimal jobs and is indicative of a large $\nu$. We can also see this by taking limits. If $\nu$ goes to zero, the $\mathbb{P}(\gamma|\Theta)$ degenerates to a single point and therefore the difference inside the brackets would be zero. On the other hand, as $\nu$ goes to infinity, the market choice probabilities converge to a uniform distribution and the differences between expected and realized systematic utility

will be large.

### A.5.2 $\xi_\gamma$

$$\ell_{\xi_\gamma} = 0 \Rightarrow \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\mathbb{1}\{\gamma_{it}=\gamma\} - \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\mathbb{P}(\gamma|\iota_i;\Theta) = 0$$

The above expression chooses $\xi$, which enters the expression through $\mathbb{P}(\gamma|\iota_i;\Theta)$, in order to equate the fraction of job switchers observed to choose market $\gamma$ with the probability that a given job-switcher would choose $\gamma$. In otherwords, $\xi$ is identified by market choices.

### A.5.3 $\phi_{\iota\gamma}$

$$\ell_{\phi_{\iota\gamma}} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\log\omega_{it} - \log\phi_{\iota\gamma_{it}}}{\phi_{\iota\gamma_{it}}}\mathbb{1}\{\gamma_{it}=\gamma,\iota_i=\iota\} + \\ + \frac{1}{\nu}\sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\mathbb{1}\{\iota_i=\iota\}\left[\mathbb{1}\{\gamma_{it}=\gamma\} - \mathbb{P}(\gamma_{it}|\iota_i;\Theta)\right] = 0$$

The above expression is highly intuitive. It tells us that identification of $\phi_{\iota\gamma}$ comes from two sources: earnings for all workers (first term), and market choices for job-switchers (second term). The first term is minimized when $\log\phi_{\iota\gamma}$ is close to actual log-earnings $\log\omega$. The second term is minimized when the theoretical probability of a type $\iota$ job-switcher choosing a job in market $\gamma$ equals the fraction of type $\iota$ job-switchers who actually choose market $\gamma$ jobs. The relative weight of these terms in calculating the likelihood is determined by the variances of measurement error in wages and idiosyncratic shocks, $\sigma^2$ and $\nu$, respectively. Specifically, if wages are observed with considerable error (large $\sigma^2$) then we put more weight on the second term, which is identified by job changes. On the other hand, if the idiosyncratic preferences have high variance (large $\nu$), then wages are more informative than job changes.

Another thing to notice is that in cases where we observe no matches for a particular $(\iota,\gamma)$ pair, identification comes purely from the second term (because $\mathbb{1}\{\gamma_{it}=\gamma,\iota_i=\iota\}=0$ in the first term). This makes sense, because we do not observe wages for matches that do not occur. Identification based on job choices in the second term relies on the assumption of a T1EV-distributed preference parameter. This is because, in order to achieve a choice probability of zero to match the count of observed matches, $\phi_{\iota\gamma}+\xi_\gamma$ will be forced towards $-\infty$. In practice, we will do something to handle zeros because we do not want to set

$\phi_{\iota\gamma}+\xi_\gamma = -\infty$. This allows us to achieve identification of the entire $\Phi$ matrix despite sparsity in observed $(\iota, \gamma)$ matches, although identification for sparse parts of $\Phi$ relies strongly on functional form assumptions. While identification based on functional form assumptions is suboptimal, we are doing so primarily for $(\iota, \gamma)$ pairs that rarely match, so imprecise estimation of these parameters will have minimal effect on our actual results. On the other hand, moving away from non-parametric identification allows us to identify a much higher degree of productivity heterogeneity.

More technically, if an $(\iota, \gamma)$ cell has zero matches, i.e. if $\mathbb{1}\{\gamma_{it} = \gamma, \iota_i = \iota\} = 0$ for all $i, t$, then the FOC above will be reduced to $\sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\mathbb{1}\{\iota_i = \iota\}\mathbb{P}(\gamma_{it}|\iota_i;\Theta) = 0$. This implies that there is no solution to the MLE problem, as $\phi_{\iota\gamma} + \xi_\gamma$ would have to go to minus infinity to make the FOC equation zero. A potential way to handle this is to add a small positive constant inside the last FOC brackets multiplied by the indicator $\mathbb{1}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T} c_{it}\mathbb{1}\{\gamma_{it} = \gamma, \iota_i = \iota\} = 0\right\}$.

### A.5.4 $\lambda$

Note that we have dropped $\iota\gamma$ subscripts here, but the estimation would be approximately the same with the subscripts.

$$
\begin{aligned}
\ell_\lambda = 0 \Rightarrow \frac{1}{\lambda}\left(\sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}\right) - \frac{1}{1-\lambda}\left((T-1)N - \sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}\right) = 0 \\
\Rightarrow (1-\lambda)\left(\sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}\right) = \lambda\left((T-1)N - \sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}\right) \\
\Rightarrow \left(\sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}\right) = \lambda(T-1)N \\
\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} c_{it}}{(T-1)N}
\end{aligned}
$$

### A.5.5 $\sigma$

Again, we have dropped $\iota\gamma$ subscripts here, but the estimation would be approximately the same with the subscripts.

We proceed taking derivatives w.r.t. $\sigma$, knowing that $f_\omega(\omega|\Theta) = \frac{1}{\omega\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log\omega - \log\phi_{\iota\gamma}}{\sigma}\right)^2} = \frac{1}{\omega\sigma}\phi\left(\frac{\log\omega - \log\phi_{\iota\gamma}}{\sigma}\right)$ and that $\log f_\omega(\omega|\Theta) = -\log(\omega\sqrt{2\pi}) - \log\sigma - \sigma^{-2}\frac{1}{2}\left(\log\omega - \log\phi_{\iota\gamma}\right)^2$

$$\ell_\sigma = 0 \Rightarrow \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\partial \log f_\omega(\omega_{it}|\Theta)}{\partial \sigma} = 0$$
$$= -\frac{NT}{\sigma} + \sigma^{-3}\sum_{i=1}^{N}\sum_{t=1}^{T}(\log \omega_{it} - \log \phi_{\iota\gamma_{it}})^2 = 0$$
$$\Rightarrow \hat{\sigma}^2 = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\left(\log \omega_{it} - \log \hat{\phi}_{\iota\gamma_{it}}\right)^2}{NT}$$

## A.6 Measurement error

The Bartik regressions in equation 1.8.2 can be written

$$\Delta Earnings_g = \beta_0 + \beta_1 Bartik_g + \varepsilon_g$$

where

$$Bartik_g = \sum_m (Exposure_{gm} \times Shock_m)$$

The earnings variable depends only on worker classifications, $g$, however the Bartik instrument depends on both worker and job classifications, $g$ and $m$. This means that worker classification error will affect both the LHS and the RHS, while job classification error will affect only the RHS.

For simplicity, Let $Y = \{\Delta Earnings_g\}_{g=1}^{G}$, $X = \{Bartik_g\}_{g=1}^{G}$, and $U = \{\varepsilon_g\}_{g=1}^{G}$. Then our regression model is

$$Y = X\beta + U$$

However we measure both $X$ and $Y$ with additive measurement error, $V_X$ and $V_Y$. Denote our measures of $X$ and $Y$, $\tilde{X}$ and $\tilde{Y}$, respectively, where

$$\tilde{X} = X + V_X$$
$$\tilde{Y} = Y + V_Y$$

If we estimate the regression using the noisy measures $\tilde{X}$ and $\tilde{Y}$ we obtain

$$\tilde{\beta} = (\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\tilde{Y})$$

For simplicity, let's assume that $X$, $V_X$, and $V_Y$ are orthogonal to the regression error term $\varepsilon$. Asymptotically,

$$\begin{aligned}\tilde{\beta} &\xrightarrow{p} \frac{Cov(X+V_X, Y+V_Y)}{Var(X+V_X)} \\ &= \frac{Cov(X+V_X, X\beta+U+V_Y)}{Var(X+V_X)} \\ &= \frac{\beta Var(X)+Cov(X,U)+Cov(X,V_Y)+\beta Cov(X,V_X)+Cov(V_X,U)+Cov(V_X,V_Y)}{Var(X)+Var(V_X)+2Cov(X,V_X)}\end{aligned}$$

For simplicity, and because we are focusing on the problem of measurement error rather than endogenous regressors, we assume that the regression error $U$ is independent of both X and $V_X$: $U \perp\!\!\!\perp X, V_X$. This implies that $Cov(X,U) = Cov(V_X,U) = 0$ and allows us to simplify the above expression to

$$\tilde{\beta} \xrightarrow{p} \frac{\beta Var(X)+\beta Cov(X,V_X)+Cov(X,V_Y)+Cov(V_X,V_Y)}{Var(X)+Var(V_X)+2Cov(X,V_X)}$$

The true coefficient $\beta$ can be written

$$\beta = \frac{Cov(X,Y)}{Var(X)}$$

and in our application we can reasonably assume $\beta > 0 \Leftrightarrow Cov(X,Y) > 0$.

To ascertain the direction of the bias created by measurement error we compare $\tilde{\beta}$ to $\hat{\beta}$. Theoretically, the direction of the bias is ambiguous. However, we can determine the sign of the bias under different assumptions about the covariances.

The simplest assumption would be that all of the covariances involving measurement error terms are 0: $Cov(X,V_Y) = Cov(X,V_X) = Cov(V_X,V_Y) = 0$. This is equivalent to classical measurement error, giving us the familiar attenuation bias result:

$$\tilde{\beta} \xrightarrow{p} \frac{Cov(X,Y)}{Var(X)+Var(V_X)} < \hat{\beta} \xrightarrow{p} \frac{Cov(X,Y)}{Var(X)}.$$

However, we almost certainly have non-classical measurement error, so let's consider what the bias would be under more reasonable assumptions. Suppose we randomly assigned workers and jobs to groups. Then both $\tilde{X}$ and $\tilde{Y}$ would simply be equal to the overall means: $\tilde{X}_g = \bar{X}\ \forall g$ and $\tilde{Y}_g = \bar{Y}\ \forall g$. This means that for large values of $Y$, $\tilde{Y} < Y$ and similarly for $X$. This implies that $Cov(X,V_X) < 0$ and $Cov(Y,V_Y) < 0$. Combining this with the fact that $Cov(X,Y) > 0$ implies that $Cov(X,V_Y) < 0$, $Cov(Y,V_X) < 0$, and $Cov(V_X,V_Y) > 0$.

Therefore,

$$\tilde{\beta} \xrightarrow{p} \frac{\beta Var(X) + \beta\overbrace{Cov(X, V_X)}^{<0} + \overbrace{Cov(X, V_Y)}^{<0} + \overbrace{Cov(V_X, V_Y)}^{>0}}{Var(X) + \underbrace{Var(V_X)}_{>0} + \underbrace{2Cov(X, V_X)}_{<0}}$$

In this case it is theoretically ambiguous whether $\tilde{\beta} > \hat{\beta}$ or $\tilde{\beta} < \hat{\beta}$. Empirically, we consistently find that $\tilde{\beta} < \hat{\beta}$. This means that it must be the case that the terms that tend to reduce $\tilde{\beta}$ — $Var(V_X)$, $\beta Cov(X, V_X)$, and $Cov(X, V_Y)$ — must dominate the terms that increase $\tilde{\beta}$ — $Cov(V_X, V_Y)$ and $2Cov(X, V_X)$.

We demonstrate this point through a simulation. We simulate a shock as described in Section 1.8.2. We estimate a series of regressions on changes in earnings by worker type on the Bartik instrument with jobs classified by market, however in each regression we randomly misclassify some percentage of workers and jobs. We loop from 0 to 100 percent of workers misclassified in intervals of five percent, and within each loop perform the same loop from 0 to 100 percent of jobs misclassified. We present the coefficients on the Bartik instrument in Figure A.3 and the $R^2$ values in Figure A.4. $R^2$ values decline approximately monotonically with the degree of misclassification in both the worker and job dimensions, as expected. By contrast, there is much less of a coherent story with the regression coefficients. Again, this is consistent with the theoretical prediction that the effect of misclassification on regression coefficients is indeterminate.

Figure A.3: Coefficient estimates with worker and job misclassification

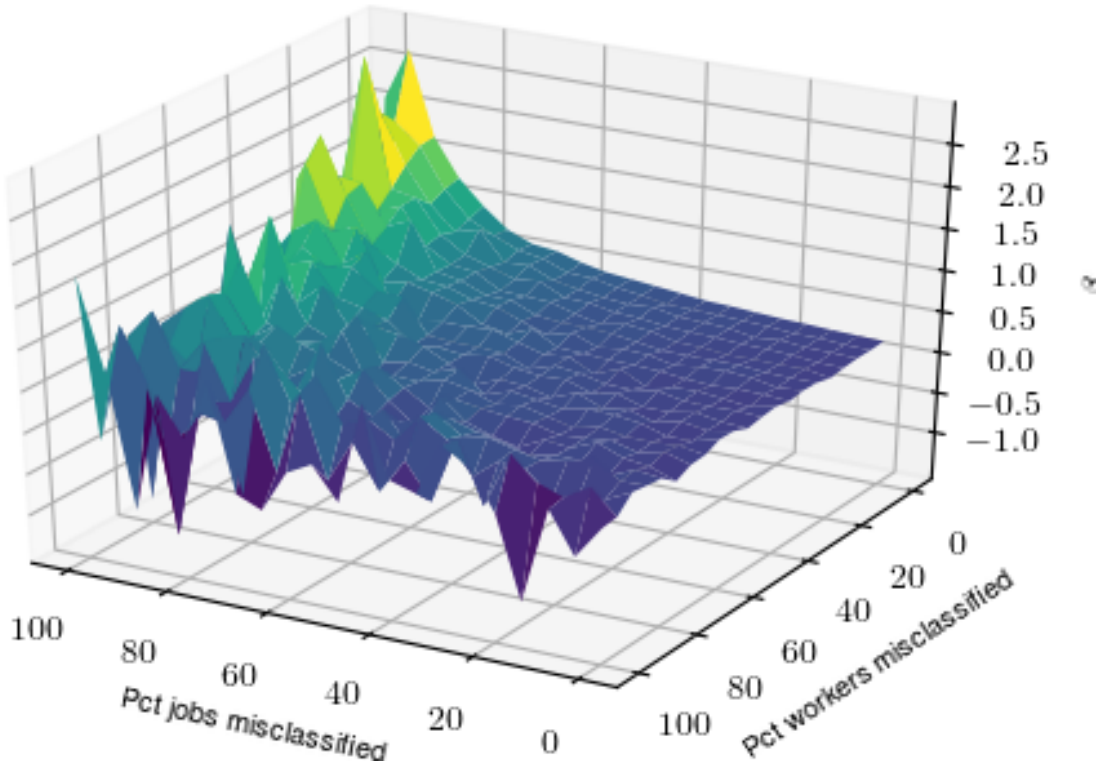
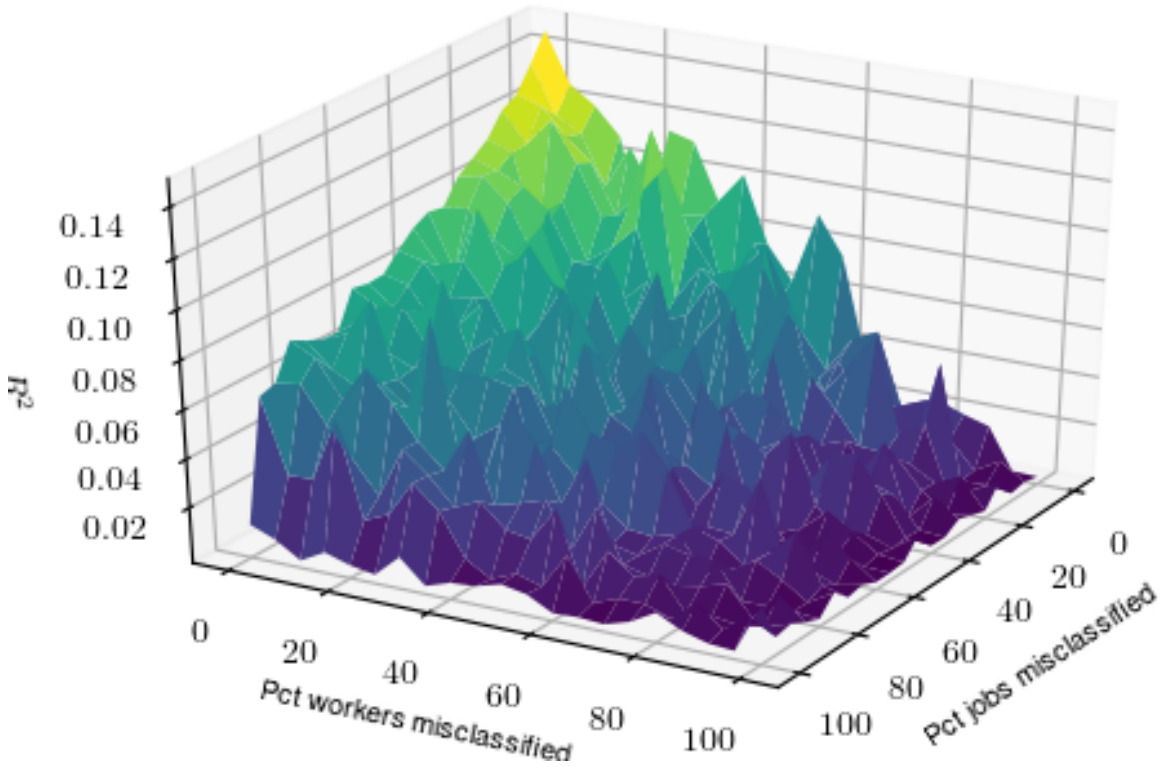


Figure A.4: $R^2$ values with worker and job misclassification



## A.7 Proof that $A_{ij}$ follows a Poisson distribution

If an individual worker $i$ only searched for a job once, then the probability of worker $i$ matching with job $j$ would be equal to $\mathbb{P}_{ij} = \mathcal{P}_{\iota\gamma} d_j$ and $A_{ij}$ would follow a Bernoulli distribution:

$$A_{ij} \sim Bernoulli(\mathcal{P}_{\iota\gamma} d_j).$$

However, since worker $i$ searches for jobs $c_i \equiv \sum_{t=1}^{T} c_{it}$ times, $A_{ij}$ is actually the sum of $c_i$ Bernoulli random variables, and is therefore a Binomial random variable. Conditional on knowing $c_i$,

$$A_{ij}|c_i \sim Binomial(c_i, \mathcal{P}_{\iota\gamma} d_j).$$

However, we still need to take into account the fact that $c_i$ is a Poisson-distributed random variable with arrival rate $d_i$. Consequently, the unconditional distribution of $A_{ij}$ is Poisson as well:

$$A_{ij} \sim Poisson(d_i d_j \mathcal{P}_{\iota\gamma}).$$

We prove this fact by multiplying the conditional density of $A_{ij}|c_i$ by the marginal density of $c_i$ to get the joint density of $A_{ij}$ and $c_i$, and then integrating out $c_i$.

$$P(A_{ij}, c_i) = \underbrace{P(A_{ij}|c_i)}_{Bin(c_i, d_j P_{\iota\gamma})} \quad \times \quad \underbrace{P(c_i)}_{Poisson(d_i)}$$

Deriving the joint distribution:

$$P(A_{ij}, c_i) = \binom{c_i}{A_{ij}} (d_j P_{\iota\gamma})^{Aij} (1 - d_j P_{\iota\gamma})^{c_i - Aij} \times \frac{d_i^{c_i} \exp(-d_i)}{c_i!}$$

We want to find out the marginal distribution of $A_{ij}$:

$$\begin{aligned}
P(A_{ij}) &= \sum_{c_i=0}^{\infty} P(A_{ij}, c_i) \\
&= \sum_{c_i=0}^{\infty} \binom{c_i}{A_{ij}} (d_j P_{\iota\gamma})^{Aij} (1 - d_j P_{\iota\gamma})^{c_i - Aij} \times \frac{d_i^{c_i} \exp(-d_i)}{c_i!} \\
&= \sum_{c_i=0}^{\infty} \frac{c_i!}{A_{ij}!(di - A_{ij})!} (d_j P_{\iota\gamma})^{Aij} (1 - d_j P_{\iota\gamma})^{c_i - Aij} \times \frac{d_i^{c_i} \exp(-d_i)}{c_i!} \\
&= \frac{(d_j P_{\iota\gamma})^{Aij} \exp(-d_i)}{A_{ij}!} \sum_{c_i=0}^{\infty} \frac{1}{(di - A_{ij})!} (1 - d_j P_{\iota\gamma})^{c_i - Aij} d_i^{c_i}
\end{aligned}$$

If the summation term is equal to

$$\sum_{c_i=0}^{\infty} \frac{1}{(di - A_{ij})!} (1 - d_j P_{\iota\gamma})^{c_i - Aij} d_i^{c_i} = d_i^{A_{ij}} \exp(d_i(1 - d_j P_{\iota\gamma})) \tag{A.1}$$

then $P(A_{ij}) = \frac{(d_i d_j P_{\iota\gamma})^{Aij} \exp(-d_i d_j P_{\iota\gamma})}{A_{ij}!}$, i.e. $A_{ij}$ would be Poisson distributed:

$$A_{ij} \sim Poisson(d_i d_j P_{\iota\gamma})$$

Proving (A.1) is equivalent to proving the following equality:

$$1 = \frac{1}{d_i^{A_{ij}} \exp(d_i(1 - d_j P_{\iota\gamma}))} \sum_{c_i=0}^{\infty} \frac{1}{(di - A_{ij})!} (1 - d_j P_{\iota\gamma})^{c_i - Aij} d_i^{c_i}$$

Proof:

$$d_i^{-A_{ij}} \exp\left(-d_i(1-d_jP_{\iota\gamma})\right) \sum_{c_i=0}^{\infty} \frac{1}{(di-A_{ij})!}(1-d_jP_{\iota\gamma})^{c_i-Aij} d_i^{c_i} =$$

$$= \sum_{c_i=0}^{\infty} \frac{\exp\left(-d_i(1-d_jP_{\iota\gamma})\right)}{(di-A_{ij})!}(1-d_jP_{\iota\gamma})^{c_i-Aij} d_i^{c_i-A_{ij}}$$

$$= \sum_{c_i=0}^{\infty} \frac{\exp\left(-d_i(1-d_jP_{\iota\gamma})\right)}{(di-A_{ij})!}(d_i(1-d_jP_{\iota\gamma}))^{c_i-Aij}$$

We assume $\lambda = d_i(1-d_jP_{\iota\gamma})$ for simplicity and we apply a change of variables $z = c_i - A_{ij}$

$= \sum_{z=0}^{\infty} \frac{\exp(-\lambda)}{z!}\lambda^z$, knowing that in our problem $c_i \geq A_{ij}$, i.e. $z \geq 0$.

$= 1$

Since we have the p.d.f. of a Poisson r.v. inside the summation, i.e. $z \sim Poisson(\lambda)$ $\square$

Therefore, we have

$$A_{ij} \sim Poisson(d_i d_j P_{\iota\gamma}) \quad \square$$

## A.8 Worker and firm fixed effects

Following Bonhomme et al. (2020) and others, we decompose the variance in workers' log earnings into a component explained by worker fixed effects, a component explained by firm fixed effects, and a component explained by the covariance between worker and firm fixed effects. We find that firm effects explain 16% of the variance in log earnings in our data and the covariance between worker and firm effects explains 11%. However, Bonhomme et al. (2020) show that estimates of the firm effects component are subject to considerable upward bias due to limited mobility of workers between firms. Therefore, building upon the approach of Bonhomme et al. (2019; 2021), we re-estimate the model at the group level, replacing firm effects with market ($\gamma$) effects. Using this grouped-data approach, we find that the share of the variance explained by market effects, as opposed to firm effects, falls to 1.2% and the share of variance explained by worker–market covariance is 2.6%.

# APPENDIX B

# Appendix to Chapter 2

## B.1 Terms in the NP decomposition

The terms in the NP decomposition from equation 2.4.2 can be more formally defined as follows:

$$\Delta_M := \left[\int_{\bar{S}_F} Y_1(x)\frac{dF_M(x)}{\mu_M(\bar{S}_F)} - \int_{S_F} Y_1(x)\frac{dF_M(x)}{\mu_M(S_F)}\right]\mu_M(\bar{S}_F) \tag{B.1}$$

$$\Delta_X := \int_{S_M\cap S_F} Y_1(x)\left[\frac{dF_M(x)}{\mu_M(S_F)} - \frac{dF_F(x)}{\mu_F(S_M)}\right]$$

$$\Delta_0 := \int_{S_M\cap S_F} [Y_1(x) - Y_0(x)]\frac{dF_F(x)}{\mu_F(S_M)}$$

$$\Delta_F := \left[\int_{S_M} Y_0(x)\frac{dF_F(x)}{\mu_F(S_M)} - \int_{\bar{S}_M} Y_0(x)\frac{dF_F(x)}{\mu_F(\bar{S}_M)}\right]\mu_F(\bar{S}_M)$$

where: $F_M(x)$ and $F_F(x)$ denote the distributions of $x$ for both males and females, respectively; $\mu_M$ and $\mu_F$ measure the proportions of males and females over regions of the supports of $x$; and the support of $x$ for a gender $g$, $supp_(X_g)$, is partitioned as $supp_(X_g) := S_g \cup \bar{S}_g$, with $S_g \cap \bar{S}_g = \emptyset$, for $g \in \{F, M\}$.

## B.2 Soft assignment workers and jobs to worker types and markets

In section 2.3, at the maximum of our posterior in equation 2.3.8, each worker is assigned to only one skill cluster, a process of *hard assignments*. However, it is possible that, given the pattern of worker matches, a particular worker could be revealed to possess certain skills $\iota_1$ in most of her matches, and skills $\iota_2$ in a few other of her matches. Creating a single worker skill group to accommodate her hybrid skills might not improve model fit if there are only a few workers who exhibit similar matches. Instead, allowing her to have mixed skills $\iota_1$ and $\iota_2$, i.e. *soft assignment*, with weights according to her matching history, provides further nuanced information to the researcher. In fact, I propose using the Bayesian setup in order to recover these weights.

It turns out that the posterior $P(\boldsymbol{b}|\boldsymbol{A}, \boldsymbol{g})$ ultimately carries the desired measure of workers' *skill profile* needed to control for workers' unobserved skills in the wage gap estimation. Given a total of $I$ clusters of workers competing for the same jobs in the labor market network, i.e. with similar skills, the posterior distribution provides the chance of each worker to belong to a certain skill cluster, given the worker demographic group $g$ and the entire network $\boldsymbol{A}$. More formally, for worker $i$, her *skills profile* is defined as:

$$\vec{P_i} := [P(i \in \iota_1|\boldsymbol{A}, \boldsymbol{g}) \qquad P(i \in \iota_2|\boldsymbol{A}, \boldsymbol{g}) \qquad \cdots \qquad P(i \in \iota_I|\boldsymbol{A}, \boldsymbol{g})]^T \tag{B.1}$$

# APPENDIX C

# Appendix to Chapter 3

Table C.1: Accuracy Scores from Predictions of 2-digit Occupation Using Random Forest Classifier (Jenks)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.3318 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.3259 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.2286 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2168 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2149 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.3088 | Yes | Yes | Yes | No | No | No | No | No |
| 0.1828 | Yes | Yes | No | No | No | No | No | No |
| 0.1831 | No | Yes | No | No | No | No | No | No |
| 0.1613 | Yes | No | No | No | No | No | No | No |
| 0.3125 | No | Yes | Yes | No | No | No | No | No |
| 0.3047 | Yes | No | Yes | No | No | No | No | No |
| 0.2999 | No | No | Yes | No | No | No | No | No |
| 0.2015 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 2-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the Jenks natural definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

Table C.2: Accuracy Scores from Predictions of 2-digit Occupation Using Random Forest Classifier (Quantile Bins)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.3256 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.3250 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.2122 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2109 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2115 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.3034 | Yes | Yes | Yes | No | No | No | No | No |
| 0.1364 | Yes | Yes | No | No | No | No | No | No |
| 0.1348 | No | Yes | No | No | No | No | No | No |
| 0.1299 | Yes | No | No | No | No | No | No | No |
| 0.3045 | No | Yes | Yes | No | No | No | No | No |
| 0.3027 | Yes | No | Yes | No | No | No | No | No |
| 0.2997 | No | No | Yes | No | No | No | No | No |
| 0.2013 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 2-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the quantile bins definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

Table C.3: Accuracy Scores from Predictions of 2-digit Occupation Using Random Forest Classifier (SEIN)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.3357 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.3255 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.2509 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2334 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.2228 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.3120 | Yes | Yes | Yes | No | No | No | No | No |
| 0.2540 | Yes | Yes | No | No | No | No | No | No |
| 0.2656 | No | Yes | No | No | No | No | No | No |
| 0.2008 | Yes | No | No | No | No | No | No | No |
| 0.3153 | No | Yes | Yes | No | No | No | No | No |
| 0.3081 | Yes | No | Yes | No | No | No | No | No |
| 0.3000 | No | No | Yes | No | No | No | No | No |
| 0.2018 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 2-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the SEIN definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

Table C.4: Accuracy Scores from Predictions of 4-digit Occupation Using Random Forest Classifier (Jenks)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.14720 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.13550 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.09059 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08486 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08344 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.12780 | Yes | Yes | Yes | No | No | No | No | No |
| 0.06710 | Yes | Yes | No | No | No | No | No | No |
| 0.07076 | No | Yes | No | No | No | No | No | No |
| 0.05790 | Yes | No | No | No | No | No | No | No |
| 0.13100 | No | Yes | Yes | No | No | No | No | No |
| 0.12050 | Yes | No | Yes | No | No | No | No | No |
| 0.11530 | No | No | Yes | No | No | No | No | No |
| 0.07599 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 4-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the Jenks natural breaks definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

Table C.5: Accuracy Scores from Predictions of 4-digit Occupation Using Random Forest Classifier (Quantile Bins)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.14070 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.13580 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.08115 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08065 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08042 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.11880 | Yes | Yes | Yes | No | No | No | No | No |
| 0.04864 | Yes | Yes | No | No | No | No | No | No |
| 0.04792 | No | Yes | No | No | No | No | No | No |
| 0.04504 | Yes | No | No | No | No | No | No | No |
| 0.12060 | No | Yes | Yes | No | No | No | No | No |
| 0.11890 | Yes | No | Yes | No | No | No | No | No |
| 0.11500 | No | No | Yes | No | No | No | No | No |
| 0.07588 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 4-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the quantile bins definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

Table C.6: Accuracy Scores from Predictions of 4-digit Occupation Using Random Forest Classifier (SEIN)

| Accuracy | Predictors | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\iota$ | $\gamma$ | NAICS3 | Earnings | Tenure | Sex | Race | Ethnicity |
| 0.15220 | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.13520 | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| 0.10450 | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08901 | No | Yes | No | Yes | Yes | Yes | Yes | Yes |
| 0.08643 | Yes | No | No | Yes | Yes | Yes | Yes | Yes |
| 0.13660 | Yes | Yes | Yes | No | No | No | No | No |
| 0.10740 | Yes | Yes | No | No | No | No | No | No |
| 0.11480 | No | Yes | No | No | No | No | No | No |
| 0.08008 | Yes | No | No | No | No | No | No | No |
| 0.14080 | No | Yes | Yes | No | No | No | No | No |
| 0.12860 | Yes | No | Yes | No | No | No | No | No |
| 0.11510 | No | No | Yes | No | No | No | No | No |
| 0.07629 | No | No | No | Yes | Yes | Yes | Yes | Yes |

*Notes:* Table presents accuracy scores from predictions of 4-digit occupation using the random forest classifier with various sets of predictors. Worker types and markets identified using the SEIN definition of jobs. The accuracy score is defined as the number of correct predictions divided by the total number of predictions: Accuracy $= \frac{\text{Num. Correct Predictions}}{\text{Num. Correct Predictions+Num. Incorrect Predictions}}$.

# BIBLIOGRAPHY

# BIBLIOGRAPHY


Acemoglu, D. and D. Autor (2011): "Skills, tasks and technologies: Implications for employment and earnings," 4, 1043–1171.

Arnold, D. (2020): "Mergers and acquisitions, local labor market concentration, and worker outcomes," *Job Market Paper. https://scholar. princeton. edu/sites/default/files/dharnold/files/jmp. pdf.*

Autor, D. H. (2013): "The 'task approach' to labor markets: an overview," .

Autor, D. H., D. Dorn, and G. H. Hanson (2013): "The China syndrome: Local labor market effects of import competition in the United States," *The American Economic Review*, 103, 2121–2168.

Autor, D. H., D. Dorn, G. H. Hanson, and J. Song (2014): "Trade adjustment: Worker-level evidence," *The Quarterly Journal of Economics*, 129, 1799–1860.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Azar, J., I. Marinescu, and M. Steinbaum (2019): "Measuring Labor Market Power Two Ways," .

Azar, J. A., I. Marinescu, M. I. Steinbaum, and B. Taska (2018): "Concentration in US Labor Markets: Evidence From Online Vacancy Data," Working Paper 24395, National Bureau of Economic Research.

Barsky, R., J. Bound, K. K. Charles, and J. P. Lupton (2002): "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," *Journal of the American Statistical Association*, 97, 663–673.

Bartik, T. J. (1991): "Who benefits from state and local economic development policies?" .

Benmelech, E., N. Bergman, and H. Kim (2018): "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" Working Paper 24307, National Bureau of Economic Research.

Blanchard, O. J. and L. F. Katz (1992): "Regional Evolutions," *Brookings Papers on Economic Activity*, 1, 1–75.

Blinder, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *The Journal of Human Resources*, 8, 436–455.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, T. L. E. Manresa, M. Mogstad, and B. Setzler (2020): "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?" .

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

——— (2021): "Discretizing Unobserved Heterogeneity," *arXiv preprint arXiv:2102.02124*.

Bound, J. and H. J. Holzer (2000): "Demand shifts, population adjustments, and labor market outcomes during the 1980s," *Journal of labor Economics*, 18, 20–54.

Broda, C. and D. E. Weinstein (2006): "Globalization and the Gains from Variety," *The Quarterly journal of economics*, 121, 541–585.

Burstein, A., E. Morales, and J. Vogel (2019): "Changes in between-group inequality: computers, occupations, and international trade," *American Economic Journal: Macroeconomics*, 11, 348–400.

Caliendo, L., M. Dvorkin, and F. Parro (2019): "Trade and labor market dynamics: General equilibrium analysis of the china trade shock," *Econometrica*, 87, 741–835.

Card, D. (1990): "The impact of the Mariel boatlift on the Miami labor market," *ILR Review*, 43, 245–257.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018): "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 36, S13–S70.

Card, D., A. R. Cardoso, and P. Kline (2015): " Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *," *The Quarterly Journal of Economics*, 131, 633–686.

Chernozhukov, V., I. Fernndez-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

Cramér, H. (1946): *Mathematical methods of statistics*, Princeton University.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

Engbom, N., G. Gonzaga, C. Moser, and R. Olivieri (2021): "Earnings Inequality and Dynamics in the Presence of Informality: The Case of Brazil," .

Firpo, S. P., N. M. Fortin, and T. Lemieux (2018): "Decomposing Wage Distributions Using Recentered Influence Function Regressions," *Econometrics*, 6, 1–40.

Fogel, J. and B. Modenesi (2022): "What is a Labor Market? Classifying Workers and Jobs Using Network Theory," .

Fortin, N., T. Lemieux, and S. Firpo (2011): "Chapter 1 - Decomposition Methods in Economics," Elsevier, vol. 4 of *Handbook of Labor Economics*, 1–102.

Frank, M. R., D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Feldman, M. Groh, J. Lobo, E. Moro, et al. (2019): "Toward understanding the impact of artificial intelligence on labor," *Proceedings of the National Academy of Sciences*, 116, 6531–6539.

Galle, S., A. Rodriguez-Clare, and M. Yi (2017): "Slicing the pie: Quantifying the aggregate and distributional effects of trade," Tech. rep., National Bureau of Economic Research.

Garcia, L. M., H. . Nopo, and P. Salardi (2009): "Gender and Racial Wage Gaps in Brazil 1996-2006: Evidence Using a Matching Comparisons Approach," Research Department Publications 4626, Inter-American Development Bank, Research Department.

Gerard, F., L. Lagos, E. Severnini, and D. Card (2018): "Assortative Matching or Exclusionary Hiring? The Impact of Firm Policies on Racial Wage Differences in Brazil," Working Paper 25176, National Bureau of Economic Research.

Gerlach, M., T. P. Peixoto, and E. G. Altmann (2018): "A network approach to topic models," *Science advances*, 4, eaaq1360.

Goldin, C. (2014): "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104, 1091–1119.

Grigsby, J. (2019): "Skill Heterogeneity and Aggregate Labor Market Dynamics," .

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27–61.

Hurst, E., Y. Rubinstein, and K. Shimizu (2021): "Task-Based Discrimination," Working Paper 29022, National Bureau of Economic Research.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2019): "Granular search, market structure, and wages," Tech. rep., National Bureau of Economic Research.

Jenks, G. F. (1967): "The data model concept in statistical mapping," *International yearbook of cartography*, 7, 186–190.

Kantenga, K. (2018): "The effect of job-polarizing skill demands on the US wage structure," .

Karrer, B. and M. E. Newman (2011): "Stochastic blockmodels and community structure in networks," *Physical review E*, 83, 016107.

Kim, R. and J. Vogel (2021): "Trade shocks and labor market adjustment," *American Economic Review: Insights*, 3, 115–30.

Larremore, D. B., A. Clauset, and A. Z. Jacobs (2014): "Efficiently inferring community structure in bipartite networks," *Physical Review E*, 90, 012805.

Lindenlaub, I. (2017): "Sorting multidimensional types: Theory and application," *The Review of Economic Studies*, 84, 718–789.

Lipsius, B. (2018): "Labor market concentration does not explain the falling labor share," *Available at SSRN 3279007*.

Maciente, A. (2013): "The determinants of agglomeration in Brazil: input-output, labor and knowledge externalities," Ph.D. thesis, University of Illinois at Urbana-Champaign.

Mahajan, A. (2006): "Identification and estimation of regression models with misclassification," *Econometrica*, 74, 631–665.

Mansfield, R. K. (2019): "How Local Are US Labor Markets?: Using an Assignment Model to Forecast the Geographic and Skill Incidence of Local Labor Demand Shocks," .

Modenesi, B. (2022): "Detailed distribution decomposition relaxing the common support assumption," .

Morello, T. and J. Anjolim (2021): "Gender wage discrimination in Brazil from 1996 to 2015: A matching analysis," *EconomiA*.

Newman, M. (2018): *Networks*, OUP Oxford.

Nimczik, J. S. (2018): "Job Mobility Networks and Endogenous Labor Markets," .

Nopo, H. . (2008): "Matching as a Tool to Decompose Wage Gaps," *The Review of Economics and Statistics*, 90, 290–299.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

Peixoto, T. P. (2013): "Parsimonious module inference in large networks," *Physical review letters*, 110, 148701.

——— (2014a): "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," *Physical Review E*, 89, 012804.

——— (2014b): "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks," *Phys. Rev. X*, 4, 011047.

——— (2017): "Nonparametric Bayesian inference of the microcanonical stochastic block model," *Physical Review E*, 95, 012317.

——— (2018): "Nonparametric weighted stochastic block models," *Phys. Rev. E*, 97, 012306.

——— (2019): "Bayesian stochastic blockmodeling," *Advances in network clustering and blockmodeling*, 289–332.

Rinz, K. (2018): "Labor Market Concentration, Earnings Inequality, and Earnings Mobility," *CARRA Working Paper Series*, 2018.

Rosvall, M. and C. T. Bergstrom (2007): "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, 104, 7327–7331.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings," *Oxford economic papers*, 3, 135–146.

Schmutte, I. M. (2014): "Free to Move? A Network Analytic Approach for Learning the Limits to Job Mobility," *Labour Economics*, 29, 49 – 61.

Schubert, G., A. Stansbury, and B. Taska (2020): "Monopsony and outside options," *Available at SSRN*.

Sorkin, I. (2018): "Ranking firms using revealed preference," *The quarterly journal of economics*, 133, 1331–1393.

Tan, J. (2018): "Multidimensional heterogeneity and matching in a frictional labor market - An application to polarization," .

Yagan, D. (2017): "Employment Hysteresis from the Great Recession," Tech. rep., National Bureau of Economic Research.