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

Jamie Fogel and Bernardo Modenesi

04/28/2023

For the latest version please click here.

Abstract

This paper 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. We microfound this tool using a standard equilibrium model of workers matching with jobs according to comparative advantage. Our classifications identify important dimensions of worker and job heterogeneity that standard classifications based on occupations and sectors miss. The equilibrium model based on our 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, we show 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 our classifications as opposed to occupations and sectors.

*Fogel Opportunity Insights, jamiefogel@g.harvard.edu. Modenesi: University of Michigan, bmodene@umich.edu. 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. This work is done in partnership with the Brazilian Institute of Applied Economic Research (IPEA). We thank John Bound, Abigail Jacobs, Matthew Shapiro, Mel Stephens, and Sebastian Sotelo for advice and guidance throughout this project. We also thank Charlie Brown, Zach Brown, Raj Chetty, Ying Fan, John Friedman, Florian Gunsilius, Nathan Hendren, Dhiren Patki, Rafael Pereira, Matthew Staiger, Dyanne Vaught, and Jean-Gabriel Young for helpful comments and discussions. We also received helpful feedback from seminar participants at the University of Michigan, Labo(u)r Day, the Urban Economics Association, Networks 2021, Yale University, Duke University, the Federal Reserve Bank of Boston, Opportunity Insights, and JAM.

1 Introduction

Many questions in economics require 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] 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.[2] This paper develops a new model-consistent approach to classifying heterogeneous workers and jobs into discrete groups based similarities revealed by observed worker-job matching patterns. Our method yields classifications that capture intuitive patterns at a high level of granularity, and in an empirical application we demonstrate that using our classifications, as opposed to traditional classifications, produces significantly larger estimates of the effects of exposure to shocks on workers' earnings.

We employ a revealed preference approach that relies on workers' and jobs' choices to classify workers and jobs. 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. This allows us to avoid relying on assumptions or our own intuition in deciding which groups of workers and jobs belong in the same clusters. For example, our method places physical education teachers and youth sports coaches in one cluster, and math and English teachers in another; had we relied on the occupation classification system, physical education teachers and youth sports coaches would have been in different classifications, even at the 1-digit occupation code level.

We formalize this intuition and apply it to large-scale data using a **?** 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."[3] Workers match with jobs according to comparative advantage, which is determined

[1] For example, **?** group workers by commuting zone, and **?** groups workers by race and predicted earnings quartile.

[2] 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 **?** 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."

[3] "Skills" and "tasks" should be interpreted broadly as any worker and job characteristics that determine

by complementarities between skills and tasks at the worker type–market level. The model implies that all workers in the same worker type have the same vector of match probabilities over jobs, and all jobs in the same market have the same vector of hiring probabilities over workers. We invert this logic and derive a maximum likelihood estimator that assigns workers to worker types and jobs to markets in the way that is most consistent with observed worker-job matches. We show that our MLE is equivalent to a tool from the community detection branch of network theory called the bipartite stochastic block model (BiSBM), and use computational techniques from network theory to solve the model.

The key to our approach is the insight that the millions of worker-job matches contained in linked employer-employee data sets can be used to estimate the worker-job match probability distribution. 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, our MLE procedure implicitly 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, respectively.

Once we have assigned workers to worker types and jobs to markets, we estimate the parameters of the labor supply Roy model. The key parameter of the model is a matrix defining the productivity of each worker type when employed in each market. We estimate 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 therefore estimate our model of the labor market as a two-step MLE procedure: the first step uses worker-job matching patterns to assign workers and jobs to worker types and markets, respectively, while the second step takes these assignments as given and estimates the underlying productivity parameters that govern comparative advantage at the worker type-market level. Finally, we embed our labor market Roy model in a calibrated general equilibrium model with workers, firms, households and exogenous product demand shocks, which propagate through the model to generate labor demand shocks. We use the general equilibrium model to simulate counterfactuals in empirical applications.

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

which workers match with which jobs.

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. However, extending our model to incorporate geographic dimensions of worker and job heterogeneity is straightforward. 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 to characterize worker and job heterogeneity.[4]

We identify worker and job types that have a high degree of granularity (290 worker types and 427 markets within the Rio de Janeiro metro area) and capture intuitive patterns that traditional classifications miss. As noted above, we aggregate workers in different occupations according to observed matching patterns (e.g. physical education teachers with youth sports coaches and math teachers with English teachers). At the same time, we *disaggregate* dissimilar workers who are employed in the same occupation. For example, among workers employed in the occupation "Course Instructor," we infer that some workers are more like coaches and physical education teachers, while others belong with math and English teachers. Additionally, 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. Finally, we demonstrate that worker types' labor supply is more concentrated within our markets than within industries (at varying levels of granularity), indicating that the markets we identify outperform industries in terms of identifying groups of jobs that are similar from the perspective of workers.

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

[4] **?** and **?** 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.

using our classifications significantly increases the magnitude of 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, first using our network-based classifications, and again 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 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.

While this paper applies our classifications to estimating the effects of local labor market shocks, they are useful for a variety of applications. For example, 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. Our classifications may be used in place of conventional classifications based on occupations, educational attainment, or low-dimensional measures of skills and tasks as a foundation in this class of models.[5] Similarly, our classifications may be used to improve labor market definitions when measuring labor market power.[6] Finally, an extension of our model to incorporate geography will allow researchers to better understand the relationship between skills/tasks and geography in determining the scope of labor markets.

Literature: We contribute to the large literature measuring the effects of labor market shocks on workers using either reduced form methods (**???????**), or a structural approach (**????**). 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 **?** and **?**, however our key innovation is identifying worker tpyes in a data-driven way and with considerable greater granularity. Our method for clustering workers and jobs builds upon the bipartite stochastic block model from the community detection branch of the network theory literature (**??**). A major contribution

[5] **??????**

[6] **??????????**

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 **?**, **?**, and **?**, 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. **?**, and **?** also use 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. **?** 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 (**??????**). 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. **?** 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 2 lays out our economic model. Section 3 builds upon the model to derive a maximum likelihood procedure for clustering workers into worker types and jobs into markets. Section 4 derives a maximum likelihood estimator for labor supply parameters, including a matrix of worker type–market match productivities. Section 5 discusses our data and sample restrictions. Section 6 presents summary statistics from our worker and job classification method. Section 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 **??** applies our classifications to Bartik-style regressions and shows that standard methods may be understating the effects of shocks on workers. Section **??** concludes.

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 5.

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 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 ι, and Γ markets, indexed by γ.
- The key parameter of the model is an $I \times \Gamma$ productivity matrix, Ψ, where the (ι, γ) cell, $\psi_{\iota\gamma}$ denotes the number of efficiency units of labor a type ι worker can supply to a job in market γ.[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.

[7] We can think of $\psi_{\iota\gamma}$ as $\psi_{\iota\gamma} = f(X_\iota, Y_\gamma)$, where X_ι is an arbitrarily high dimensional vector of skills for type ι workers, Y_γ is an arbitrarily high dimensional vector of tasks for jobs in market γ, and $f()$ is a function mapping skills and tasks into productivity. This framework is consistent with **?**'s skill and task-based model, and is equivalent to **?** and **?**. A key difference is that **?** and **?** 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.

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

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

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

$$y_s^D = \frac{a_s}{\sum_{s'} \left(\frac{p_s}{p_s'} \right)^{\eta} (a_{s'} p_{s'})} Y \tag{2}$$

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: $Y = \bar{W} + \Pi$.

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 γ. The quantity of the sector s good produced by firm f, y_{sf}, is therefore given by

$$y_{sf} = \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} \tag{3}$$

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

equation (3). Therefore, the firm solves

$$\pi_f = \max_{\{\ell_{\gamma f}\}_{\gamma=1}^{\Gamma}} \quad p_s \cdot \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} - \sum_{\gamma} w_\gamma \ell_{\gamma f}. \tag{4}$$

Production exhibits decreasing returns to scale because

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

where α denotes the labor share.

We define a job, indexed by j, as a firm-market pair. Therefore, we can replace the γf 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: $\Pi = \sum_{s=1}^{S} \sum_{f \in s} \pi_f$.

2.4 Workers

Workers, indexed by i, are endowed with a *worker type*, indexed by ι, and one indivisible unit of labor. We denote worker i's type as $\iota(i)$. There is an exogenously-determined mass of type ι workers, m_ι. The worker's type defines their skills. Type ι workers can supply $\psi_{\iota\gamma}$ efficiency units of labor to jobs in market γ. $\psi_{\iota\gamma}$ is a reduced form representation of the skill level of a type ι worker in the various tasks required by a job in market γ. 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 γ (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_γ. Therefore, the law of one price holds for each market, and a type ι worker employed in a job in market γ 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_γ.[8] In addition to earnings, each market γ has a fixed amenity value to workers, ξ_γ; $\Xi = \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_\Gamma \end{bmatrix}$. 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,

[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 concurrent work (**?**).

which is normalized to 0 without loss of generality. Finally, each worker i has an idiosyncratic preference for market γ jobs at time t, $\varepsilon_{i\gamma t}$. Therefore, worker i chooses a market by solving

$$\gamma_{it} = \underset{\gamma \in \{0,1,\ldots,\Gamma\}}{\arg\max}\ \psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma + \varepsilon_{i\gamma t} \tag{5}$$

where γ_{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 ν:

Assumption 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 ν and are serially uncorrelated and independent of all other variables in the model.*

This gives us a functional form for the probability that a type ι worker chooses a job in market γ:

$$\mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w_t}, \Xi, \nu] = \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)}. \tag{6}$$

We aggregate over individual workers to specify labor supply. As noted above, m_ι denotes the exogenously-determined mass of type ι workers. The *number* of workers employed in market γ 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 γ jobs must equal demand. To get the *quantity* of labor supplied to market γ jobs, rather than the number of workers, we weight the equation above by the number of efficiency units of labor supplied by a type ι worker to a job in market γ: $\psi_{\iota\gamma}$:

$$LS_\gamma(\vec{w_t}) = \sum_\iota m_\iota \mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w_t}, \Xi, \nu]\psi_{\iota\gamma} = \sum_\iota m_\iota \left(\frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma} \tag{7}$$

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 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 (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 2.3).

Assumption 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 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 γ_{it} according to $\mathbb{P}_\iota[\gamma_{it}|\vec{w}]$
4. Separated workers randomly match with a job within their chosen market γ

Assumptions 2.2 and 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 3.3.

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 (4)
2. Labor supply is consistent with workers' expected utility maximization (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: $y_s = \sum_{f \in s} \prod_\gamma \ell_{\gamma f}^{\beta_{\gamma s}}$ (3).
4. The labor market clears for each market γ: $LS_\gamma = LD_\gamma := \sum_s \sum_{f \in s} \ell_{\gamma f}$
5. Aggregate consumption is equal to income: $Y = \sum_s p_s y_s^D = \bar{W} + \Pi$.

2.7 Discussion

The matrix

$$\Psi = \begin{array}{c} \\ \iota = 1 \\ \iota = 2 \\ \vdots \\ \iota = I \end{array} \begin{array}{c} \begin{array}{cccc} \gamma = 1 & \gamma = 2 & \cdots & \gamma = \Gamma \end{array} \\ \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1\Gamma} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2\Gamma} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{I1} & \psi_{I2} & \cdots & \psi_{I\Gamma} \end{pmatrix} \end{array} \tag{8}$$

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 Ψ, $\psi_{\iota\gamma}$, captures the effective units of labor a type ι worker can supply to a job in market γ. Therefore, Ψ governs both absolute and comparative advantage. Each row of Ψ, $\psi_\iota = \begin{bmatrix} \psi_{\iota 1} & \psi_{\iota 2} & \cdots, & \psi_{\iota\Gamma} \end{bmatrix}$, represents a productivity vector for type ι workers and is a reduced form representation of their skills.

Ψ embeds a flexible notion of skills. It allows us to say that a particular type of worker is highly skilled *in market* γ, 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.

Ψ 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 Ψ 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

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

3 Classifying workers and jobs

In this section we derive our procedure for assigning workers to worker types, ι, and jobs to markets, γ, 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, ι, if they have the same vectors of match probabilities over markets, and two jobs belong to the same market, γ, if they have the same vectors of match probabilities over worker types.

3.1 Assigning workers to worker types and jobs to markets

As stated in equation (6), when any worker i belonging to type ι searches for a job, the probability that they choose a job in market γ 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 (Ψ, Ξ, and ν) 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 (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 γ, not a particular job j. To obtain the probability that worker i matches with a *specific* job j in market γ, we multiply the choice probability in equation (6) by the probability that worker i matches with job j, conditional on choosing a job in market γ. 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 γ 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 γ employment can be written

$$\mathbb{P}[j|\gamma] = d_j / \sum_{j' \in \gamma} d^J_{j'}. \tag{9}$$

Therefore, when worker i of type ι searches, the probability that the search results in worker i matched with job j is the product of the probabilities in equation (6) and equation (9):

$$\mathbb{P}_{ij} = \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)}}^{\mathbb{P}_i[\gamma|\Psi,\vec{w},\Xi,\nu]} \times \overbrace{\underbrace{\frac{1}{\sum_{j' \in \gamma} d^J_{j'}}}_{\substack{1/\text{ type } \gamma \\ \text{employment}}} \times \underbrace{d_j}_{\substack{\text{Job } j \\ \text{employment}}}}^{\mathbb{P}[j|\gamma]}. \tag{10}$$

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

The first term represents the probability that worker i chooses market γ, while the second represents the probability that worker i chooses job j conditional on choosing market γ. 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:

$$\begin{aligned}\mathbb{P}_{ij} &= \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma}w_\gamma+\xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma}\exp\left(\frac{\psi_{\iota\gamma'}w_{\gamma'}+\xi_{\gamma'}}{\nu}\right)} \times \frac{1}{\sum_{j'\in\gamma} d_{j'}^J}}^{:=\mathcal{P}_{\iota\gamma}} \times \underbrace{d_j}_{\substack{\text{Job } j\\ \text{employment}}} \\ &= \mathcal{P}_{\iota\gamma} d_j.\end{aligned} \tag{11}$$

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

$$A_{ij} \sim Poisson\left(d_i d_j \mathcal{P}_{\iota\gamma}\right). \tag{12}$$

For a complete proof, see appendix **??**.

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

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right) = \prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{13}$$

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.

$$\vec{\iota},\vec{\gamma} = \underset{\substack{\{\vec{\iota}=\iota(i)\}_{i=1}^N,\\ \{\vec{\gamma}=\gamma(j)\}_{j=1}^J}}{\arg\max} \prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{14}$$

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 (14) 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 Γ 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 (**???**). See appendix **??** for greater detail.

Equation (14) corresponds to the degree-corrected bipartite stochastic block model (BiSBM), a workhorse model in the community detection branch of network theory (see appendix **??** 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 (14), 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 **?** for details.)

3.2 Visual intuition of the BiSBM

Figure 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

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

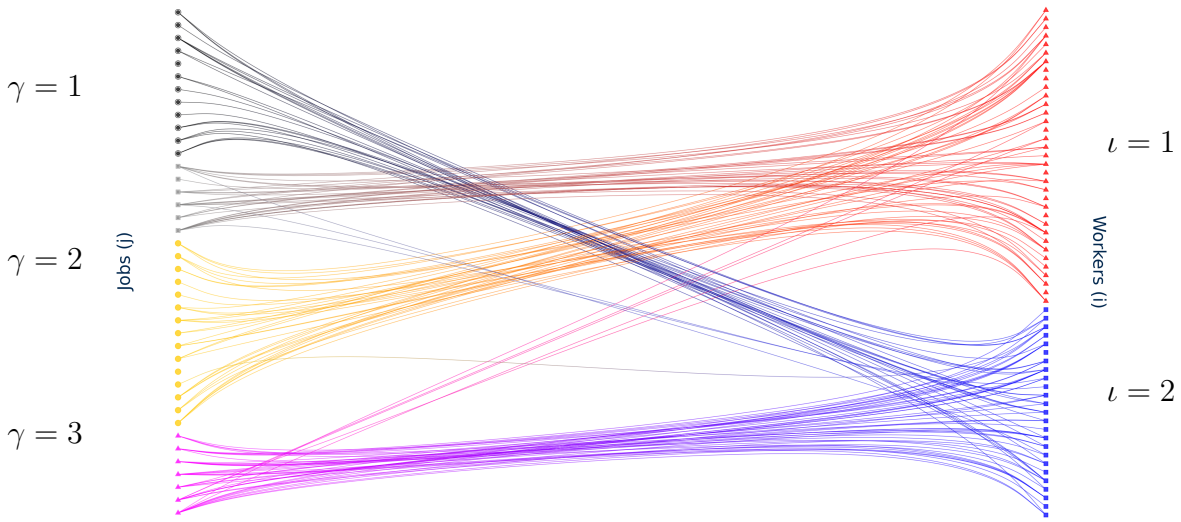
Dots on the left axis represent individual jobs j and dots on the right axis represent individual workers i. Workers belong to one of two worker types ($\iota \in \{1, 2\}$) and jobs belong to one of three markets ($\gamma \in \{1, 2, 3\}$). Lines represent employment contracts between individual workers 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. Therefore, our task, formalized in the maximum likelihood procedure defined in equation (14), 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.

3.3 Discussion

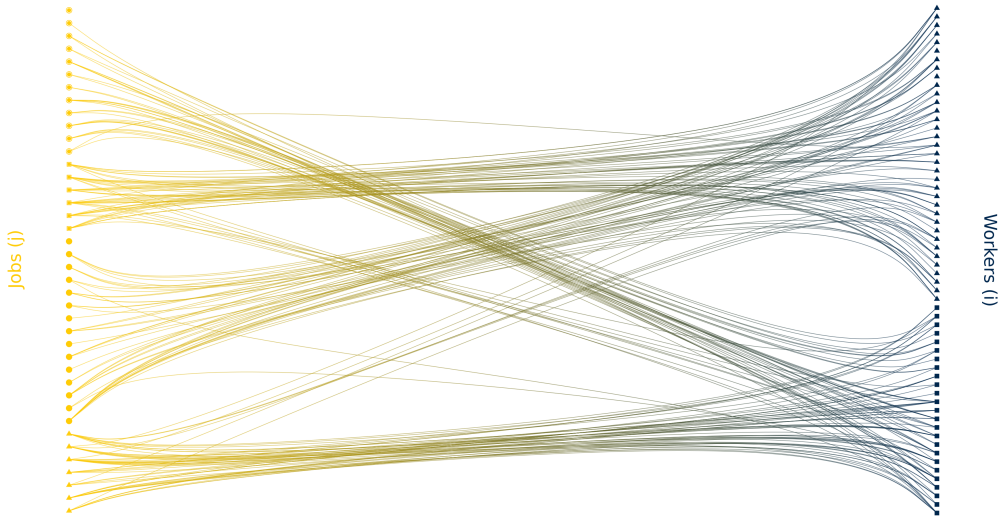
Our approach rests on the insight that workers with similar propensities to match with particular jobs have similar skills, while jobs with similar propensities to hire particular workers require similar tasks. We formalize this by making three major assumptions. First, our model implicitly assumes that 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 6). Second, Assumption 2.3 states that the fundamentals of the economy — the labor supply parameters Ψ, Ξ, and ν, and the demand shifters $\vec{a}$ — are fixed throughout our estimation window. Third, combining the assumptions of i.i.d. T1EV preference shocks (Assumption 2.1) and exogenous separations (Assumption 2.2), we assume that movement of workers between jobs represents idiosyncratic lateral moves. This

Figure 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

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

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 (13), 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 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 Ψ, Ξ, and ν, 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 ε_{it}, however exogenous separations are more tractable.[12] An implication of the exogenous separations assumption is that a worker's match probabilities are independent of their job history, conditional on their type.[13]

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.

[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 **?**, Appendix D for details on this alternative approach.

[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.

4.1 Estimating Ψ from observed matches

Identification and estimation of the labor supply parameters builds upon **?** and **?**, 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 unit mean, summarized by parameter vector θ_e. Observed earnings ω_{it} are therefore

$$\omega_{it} = \psi_{\iota(i)\gamma_{it}} w_{\gamma_{it}} e_{it}. \tag{15}$$

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

Assumption 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, ι_i, and current market choice γ_{it}.*

Our model is identified by combining assumption 4.1 with assumptions 2.1 and 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, γ_{it}, the worker's earnings, ω_{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

[14] More precisely, **?** model workers matching with firms and therefore use k-means clustering to cluster firms on the basis of the firms' earnings distributions, while **?** 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 **?** nor **?** 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.

worker i choosing a job in market γ and comes from the Roy model (equation 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 γ_{it}, and the worker's earnings ω_{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 ω_{it}.[15] Assumptions 2.2 and 4.2 tell us that realizations of ω_{it} and c_{it} are independent, and γ_{it} is independent of ω_{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

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

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:

$$\begin{aligned}\mathcal{L}(\Theta|\mathbb{X}) &= \prod_{i=1}^{N}\prod_{t=1}^{T} l(\gamma_{it}, \omega_{it}, c_{it}|\mathbb{X}) \\ &= \prod_{i=1}^{N}\prod_{t=1}^{T} \underbrace{[\mathbb{P}(\gamma_{it}|\Theta) f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{Separation}}^{c_{it}} \underbrace{[f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{No separation}}^{1-c_{it}} \end{aligned} \tag{16}$$

Finally, the log-likelihood is

$$\ell(\Theta|\mathbb{X}) = \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it} \log \mathbb{P}(\gamma_{it}|\Theta) + \sum_{i=1}^{N}\sum_{t=1}^{T} \log f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta) \tag{17}$$

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

[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.

Assumption 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 4.3 (Ψ normalization)**.** *The mean productivity level in each market γ is normalized to a constant, k:*

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

where m_ι is the mass of type ι workers.

Assumption 4.2 assumes that wages follow a log-normal distribution which is worker type-market specific, following **?** and **?**. Assumption 4.3 normalizes the $\psi_{\iota\gamma}$ to have a mean equal to some constant k within market.

Identification of Ψ 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, (ι, γ) 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 ν), estimation puts more weight on earnings. The normalization that the mean skill level in each market equals k (Assumption 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, ν, is identified by workers' choices of markets, γ. 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, ν. The market amenities parameter ξ_γ is a market fixed effect and is identified by the component of the frequency with which workers choose market γ that is common across all worker types ι. The relative value of ξ_γ 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 ξ_γ reflects not just the immediate

[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.

utility benefits of working in a job in market γ, but also reflects broader compensating differentials. In this way, ξ_{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 **??**.

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 γ) — We calibrate these parameters as the share of the sector S wage bill paid to workers employed in market γ jobs.
- η (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).

4.3 Discussion

The worker type–market productivity matrix Ψ 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 Ψ 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 Ψ. In order to do so, researchers identify clusters of similar workers and clusters of similar jobs using observable worker and job characteristics. For example, **?** imputes worker skill groups using information on workers' training and educational degrees, and defines occupation groups using skill requirement information from O*NET. Similarly, **?** 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 **?** note, "according to O*NET, the skill 'installation' is equally important to both

[17] **?** 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 7, but does not affect the reduced form estimates in Section **??**.

[18] Other papers employing a similar framework include **???**

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 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 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 6. A similar logic applies to clustering jobs into types.

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

[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.

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 (Ψ, Ξ, and ν) 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 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,

[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.

[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.

Table 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

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.

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 presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions.

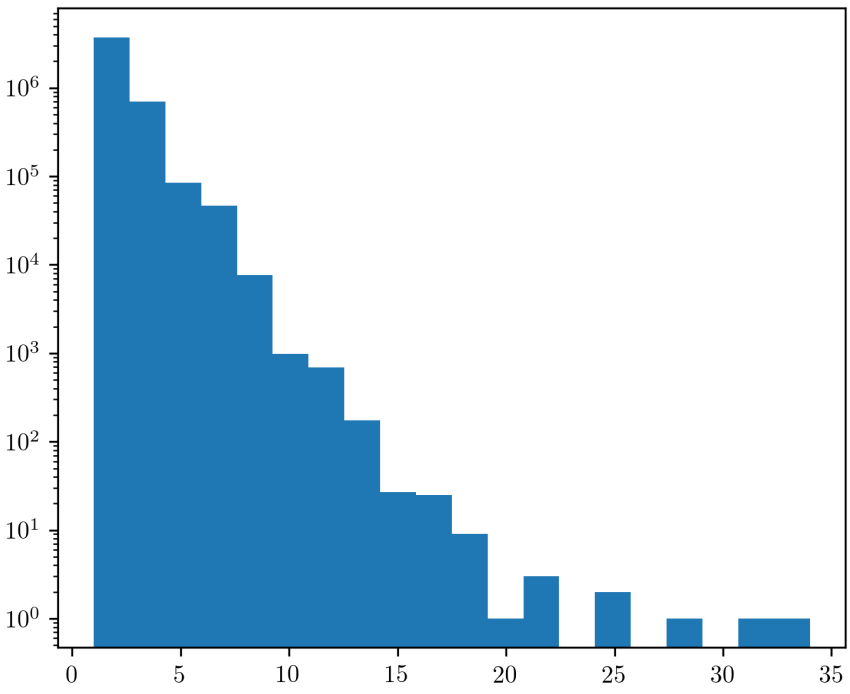
Table 2 presents the fraction of job changes that also involve a change in occupation,

[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.

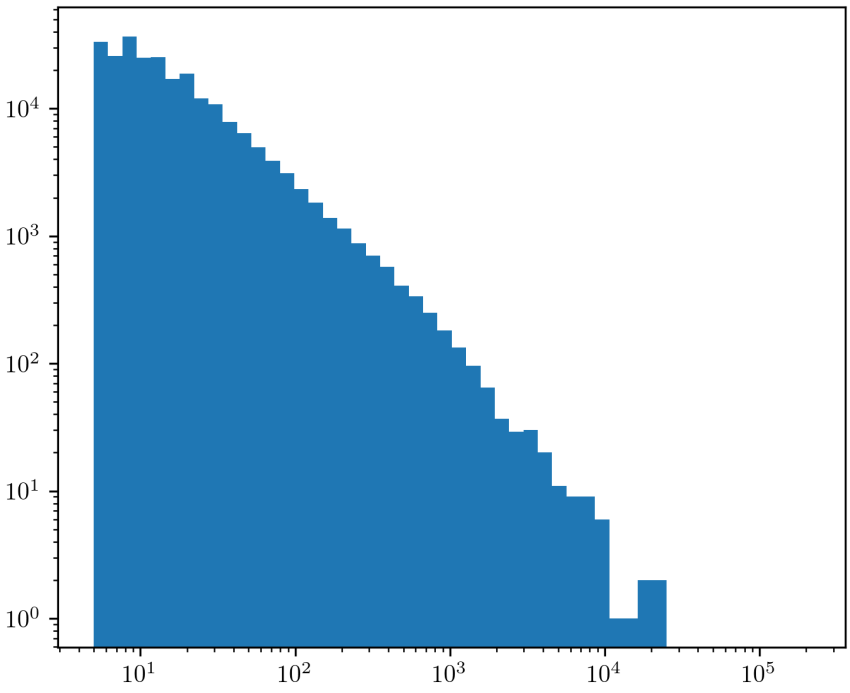
[24] See **?**, Figure 21.

Figure 2: 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 5.

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 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 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 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 (γ) 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 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 (γ)	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 × "Firm Change Only" + (1-0.536) × "No Firm Change." 5-digit sectors refer to narrow industry codes, while there are 15 IBGE sectors, defined in Table 1, taken from the Brazilian Institute of Geography and Statistics (IBGE). Values computed using the worker earnings panel described in Section 5 using RAIS data from 2009–2012.

6 Descriptive results

Our network-based classification algorithm identifies 290 worker types (ι) and 427 markets (γ). Figure 3 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.

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.

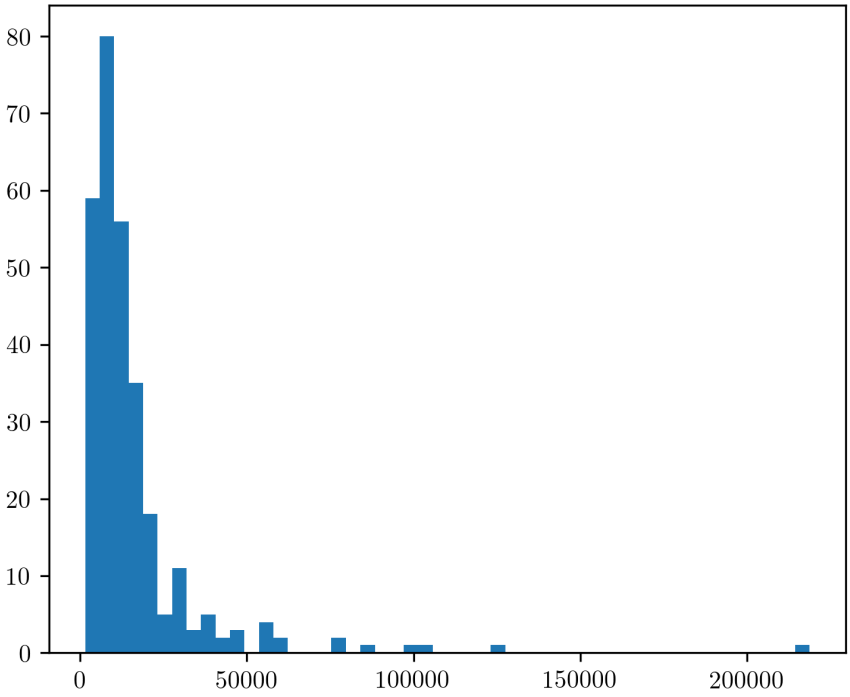
Despite this being the most disaggregated level of the occupation classification system (6-digit), there may be considerable heterogeneity within this occupation. This occupation may include, for example, both math tutors and personal fitness trainers — two sets of workers with very different skills. 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 3 and 4.

Table 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

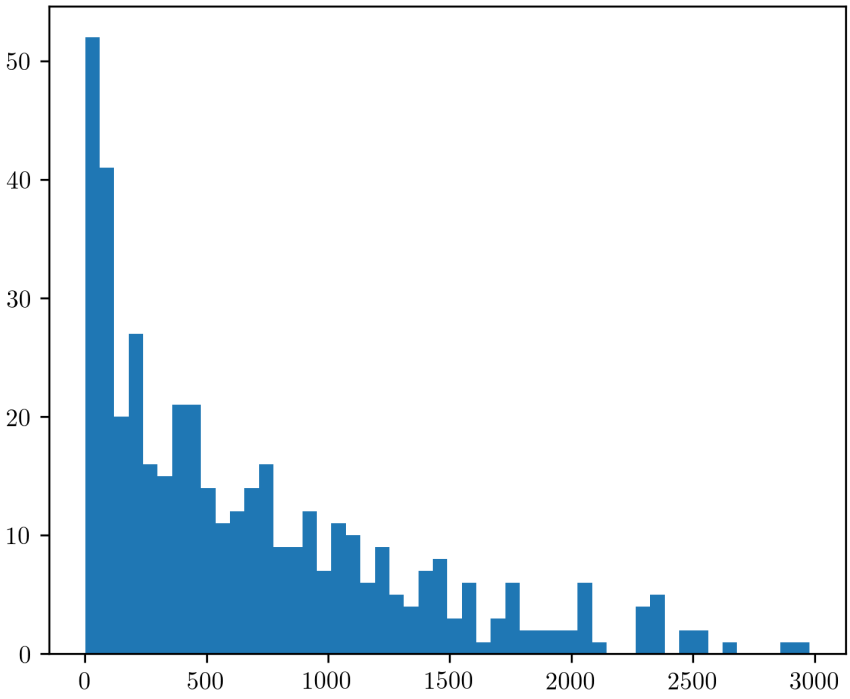
[25] Occupation names and descriptions are translated from Portuguese using Google Translate and some translations are imprecise, although manual inspection of a subset by our Portuguese-speaking coauthor confirms that most translations are satisfactory.

Figure 3: Worker Type (ι) and Market (γ) Size Distributions

(a) Number of Workers Per Worker Type (ι)



(b) Number of Jobs Per Market (γ)



Notes: Figure presents histograms of the number of workers per worker type ι and jobs per market (γ). 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 3.

individual worker to a worker type, ι. 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 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 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.

6.2 Worker type skill correlations

While Section 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

Table 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 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 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 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.

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 Ψ represents the productivity of a type ι worker employed in market γ. Therefore, $\psi_{\iota\gamma}$ is a summary measure of a type ι worker's skill at jobs in market γ, and a full row vector of Ψ, $\psi_{\iota\cdot}$, summarizes a type ι worker's skills in *all* markets. This yields a natural metric for skill similarity across worker types: two worker types, ι and ι', 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, 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 ι, ι'. 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 Ψ 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 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 290×290 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 Ψ using the same maximum likelihood estimation described in Section 4, except we classify workers and jobs by occupation and sector, rather than worker type and market. Row 2 of Figure 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 (γ). 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 4 shows that correlations between different worker types' productivity vectors are smaller in magnitude when we use our model's (ι, γ) 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.

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.

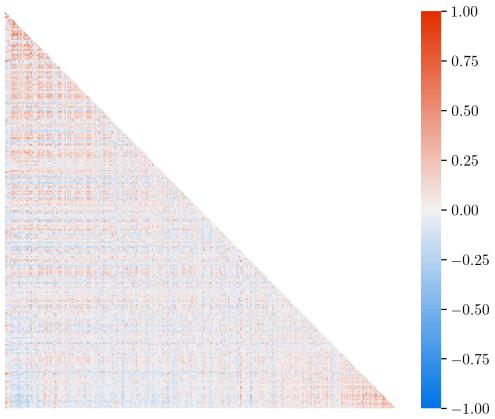
We compute each worker type's employment concentration across sectors and markets using the Herfindahl-Hirschman index (HHI):

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

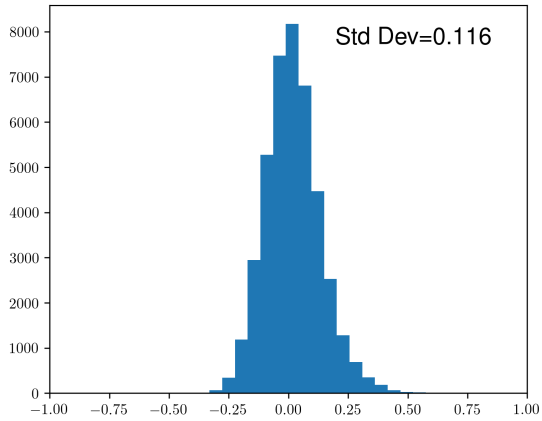
where s indexes sectors, γ indexes markets, and $\pi_{\iota s}$ and $\pi_{\iota \gamma}$ are the share of type ι workers employed in sector s and market γ, respectively. An HHI close to 0 indicates that type ι employment is spread approximately evenly across sectors/markets, while an HHI close to 1 indicates that type ι employment is very concentrated in a single sector/market. Suppose we classified jobs randomly. Then worker types would not have a comparative advantage in

Figure 4: Skill Correlation Across Worker Types and Occupations

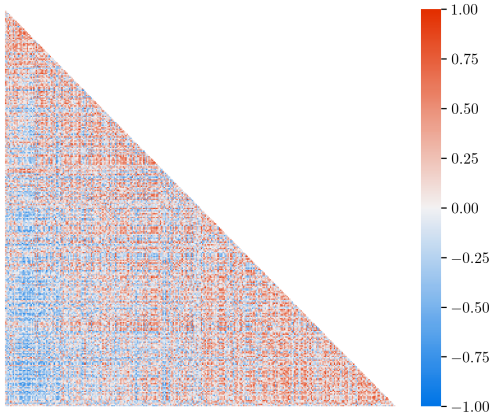
(a) (ι, γ) correlogram



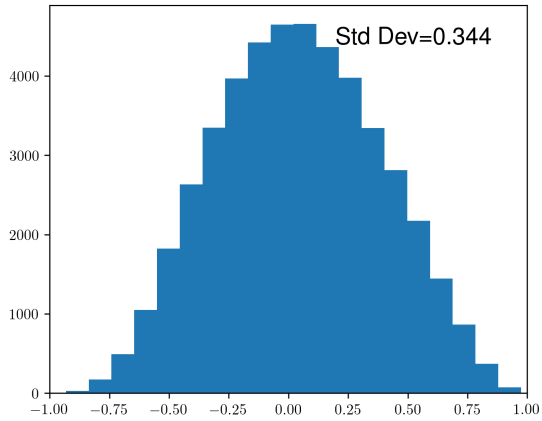
(b) (ι, γ) histogram



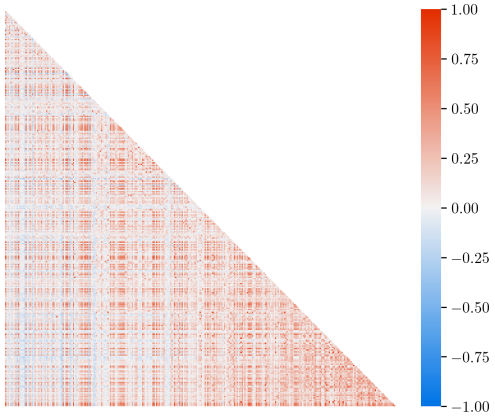
(c) (Occ4, Sector) correlogram



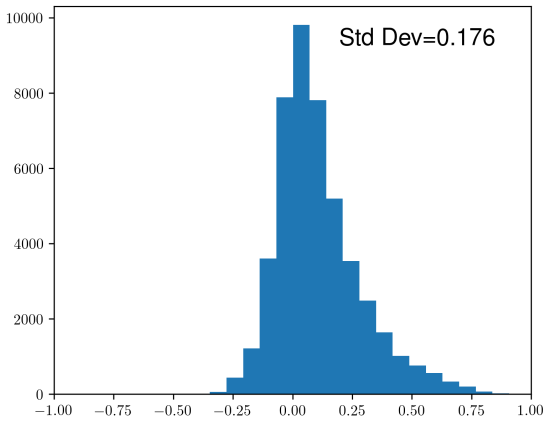
(d) (Occ4, Sector) histogram



(e) (Occ4, γ) correlogram



(f) (Occ4, γ) histogram



Notes: Figure presents pairwise skills vector correlations (left column) and histograms of these skill correlations (right column) for all pairs of worker types ι (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 Ψ, $\psi_{\iota\cdot}$, where Ψ is estimated as described in Section 4.1 using the 2009-2012 RAIS worker earnings panel described in Section 5. Workers classified by worker types ι in row 1 and by 4-digit occupation in rows 2 and 3. Jobs classified by market γ 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).

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 5a presents HHI_ι^{Sector} and HHI_ι^{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 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 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.

7 General equilibrium effects of Rio de Janeiro Olympics

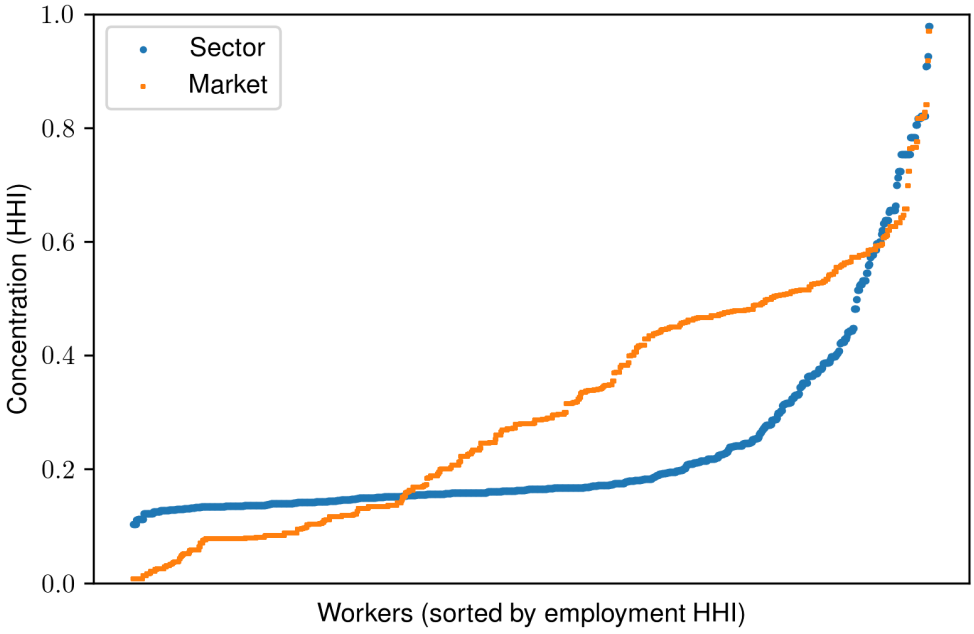
We test our model's ability to predict the effects of shocks in the context of the infrastructure investment and other preparations for the 2016 Rio de Janeiro Olympics. The Olympics were announced in late 2009 and construction of new venues and infrastructure were in full effect by 2014. Therefore, 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.

$$\Delta y = \beta_0 + \beta_1 \Delta\hat{y} + \varepsilon \tag{18}$$

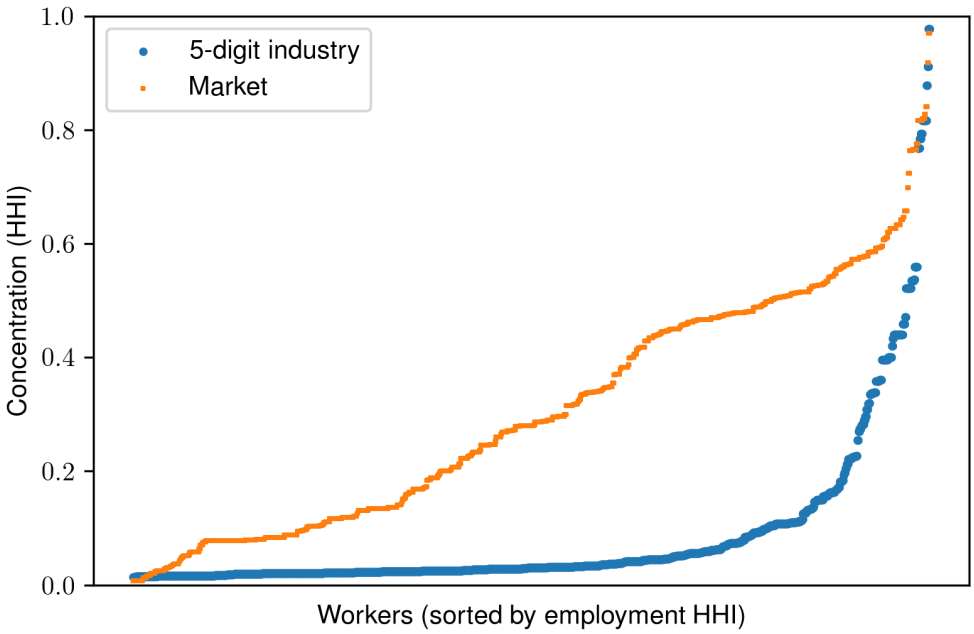
If our model is able to perfectly predict the actual effects of the Rio Olympics shock, the slope would be 1 and the intercept 0. As shown in the first column of Table 5 the slope of the best fit line is 0.982 and the intercept is -0.003, very close to our goals of 1 and 0,

Figure 5: Concentration of Worker Types' (ι) Employment Within Markets/Sectors

(a) Markets (γ) and sectors



(b) Markets (γ) 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 5.

respectively.[26]

Table 5: Predicted Effect of Olympics on Wages: Network-Based vs. Standard Classifications

Worker classification	ι	Occ4	Occ4	k-means	k-means
Market classification	γ	Sector	γ	Sector	γ
Intercept	-0.003	-0.001	-0.002	-0.0	-0.002
	(0.009)	(0.009)	(0.009)	(0.011)	(0.01)
Model-implied Δ log earnings	0.982	0.148	0.428	0.234	0.56
	(0.551)	(0.434)	(0.185)	(0.575)	(0.259)
MSE	0.021	0.025	0.025	0.023	0.023
Observations	290	306	306	214	214

Notes: Table presents results from estimating equation (18) for various worker and job classifications. Workers classified by worker type (ι) 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 (**?**). Jobs are classified by market (γ) 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 7. Dependent variables based on data from the 2009-2012 RAIS worker earnings panel described in Section 5. Independent computed by solving the model described in Section 2 using parameters estimated in Section 4.1 and calibrated in Section 4.2. Regressions are weighted by the number of workers per classification.

[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 3).