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

Jamie Fogel and Bernardo Modenesi

12/12/2023

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 study the labor demand shock generated by trade liberalization with China, Autor et al. (2013) group workers by commuting zone (CZ) and compare outcomes between workers in CZs with varying levels of exposure to the China shock. Other commonly-used indicators of worker and job similarity include observable variables like age, education, occupation, industry, geography, or skills as measured by the Occupational Information Network (O*NET). Relying on observable indicators of worker and job similarity has well-known 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. Indeed, as Autor and coauthors note, "according to O*NET, the skill 'installation' is equally important to both computer programmers and to plumbers, but, undoubtedly, workers in these occupations are performing very dissimilar tasks." (Frank et al., 2019)

To understand why this matters, consider a simplified version of Autor et al. (2013)'s China shock. Suppose there are two CZs, A and B, where we assume A was exposed to the China shock and B was not. To estimate the effect of the China shock, we would compare the difference in pre- to post-shock changes in outcomes between CZs A and B. Suppose further that (i) the truly relevant heterogeneity can be characterized as urban versus rural, (ii) the China shock only affected rural areas, and (iii) CZ A is 2/3 rural and B is 1/3 rural. In this case, when we compare changes in outcomes between A and B we are comparing outcomes for a "treatment" group that is in fact 2/3 treated (rural) and 1/3 control (urban) to outcomes for a "control" group that is actually 2/3 control and 1/3 treated. Consequently, our estimated effects may be attenuated. In general, the relevant dimensions of heterogeneity may encompass both skill/task and geographic dimensions, and cut across observable groups. Therefore, instead of relying on observable characteristics, we propose a novel, data-driven approach that classifies workers and jobs into discrete groups based on patterns of similarity revealed by the network of worker–job matches embedded in linked employer-employee data.

We start with the intuition that workers employed in the same job have similar skills, and jobs employing the same workers require similar tasks. We formalize this intuition using a Roy (1951) model in which workers belong to a discrete set of latent *worker types* and jobs belong to a discrete set of latent *markets*. For ease of exposition, we assume that worker types and markets reflect only skills and tasks, respectively, however they may also reflect factors like geography, credentials, or preferences. Workers match with jobs according to comparative advantage, which is determined 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 based on the observed network of worker–job matches. The MLE estimator uses realized job matches for each worker and their peers — coworkers, former coworkers, coworkers' former coworkers, former coworkers' coworkers, etc. — to approximate each worker's match probability distribution over jobs, and clusters together workers with the most similar distributions into worker types. It simultaneously clusters jobs into markets following a symmetric argument. 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.

By inferring worker and job similarity directly from empirical matching patterns, we bypass the need to (i) directly observe all relevant dimensions of worker and job heterogeneity, and (ii) understand how those dimensions interact in potentially non-linear ways. Moreover, by deriving our clustering algorithm from a Roy model of workers matching with jobs according to comparative advantage, we give our worker types and markets precise economic interpretability. Finally, we embed our Roy model of the labor market within a general equilibrium model with workers, firms, households and exogenous product demand shocks based on Grigsby (2019) in order to microfound the determinants of comparative advantage and allow for counterfactuals representing labor supply or demand shocks.

We estimate our model and conduct empirical analyses using Brazilian administrative records from the Annual Social Information Survey (RAIS) that is managed by the Brazilian labor ministry. The RAIS data contain detailed information about every formal sector employment contract, including worker demographic information, occupation, sector, and earnings. Critically, these data represent a network of worker–job matches in which workers are connected to every job they have ever held, and vice versa.

We identify 446 worker types and 1,371 markets. This level of granularity is similar to defining markets by the intersection of two-digit occupations and meso regions[1], henceforth "occ2Xmesos", of which there are 1,480. Therefore, we treat occ2Xmesos as our primary "status quo" comparison. We present descriptive facts about our worker types and markets that (i) validate the quality of our worker types and markets, and (ii) identify patterns of worker and job similarity that standard classifications based on occupation, industry or geography would have missed.

[1]Meso regions are administrative divisions that are larger than cities but smaller than states. There are 137 meso regions in Brazil.

Our validation exercises begin with the premise that the ideal worker and job classifications will maximize within-group similarity and minimize across-group similarity. First, we show that workers' labor supply is significantly more concentrated within our markets than within either occ2Xmesos or industries, and symmetrically jobs' hiring is more concentrated within our worker types. Second, we perform an out-of-sample prediction exercise in which our worker types and markets outperform occ2Xmesos in predicting workers' future job-to-job flows. Third, we estimate the worker type–market productivity matrix that governs comparative advantage in our Roy model, which can be interpreted as a representation of worker skills, and show that our worker types do a better job of identifying groups of workers with distinct skills. Finally, we use our general equilibrium model to show that our classifications more accurately predict the wage effects of the labor demand shock created by the 2016 Rio de Janeiro Olympics.

The clusters we identify are intuitive. For example, we identify one worker type composed primarily of physical education teachers and youth sports coaches and another composed of math and English teachers, despite the fact that these occupations differ at the two-digit level. 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. We find that some markets are defined more by skills and tasks, while others are defined more by geography. Additionally, the distribution of occupations within some markets is highly concentrated, while within others it is dispersed. Traditional market definitions based on occupation and/or geography alone would have missed important details.

Our main empirical application applies our worker types and markets to reduced form Bartik-style regressions and finds that 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, our classifications may be used to improve labor market definitions when measuring labor market power.[2] Similarly, to characterize two-sided (worker–job) multidimensional heterogeneity, researchers identify groups of workers with similar skills and study how they match with groups of jobs requiring similar tasks. Our classifications may be used in place of conventional classifications — based on occupations, educational attainment, or low-dimensional measures of skills and tasks — in this class of models.[3]

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

Conditional on assigning workers to latent worker types and jobs to latent markets, our model of labor supply is similar to Grigsby (2019) and Bonhomme et al. (2019), however our key innovation is identifying worker types 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 (Larremore et al., 2014; Peixoto, 2019). A major contribution of our paper is creating a theoretical link between a labor supply model and the BiSBM, thereby providing microfoundations for using tools from network theory to solve problems in economics and giving these tools clear economic interpretability.

Like Sorkin (2018), Nimczik (2018), and Jarosch et al. (2019), we use tools from network theory to extract previously unobserved information from LEED. We use the panel of worker–job matches to identify worker and job *similarities*; by contrast, Sorkin exploits the direction of worker flows between firms to identify *differences* between firms. Nimczik (2018), and Jarosch et al. (2019) 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. Schmutte (2014) uses a different tool from

[2]Berger et al. (2022); Felix (2021); Azar et al. (2018); Benmelech et al. (2018); Rinz (2018); Azar et al. (2019); Schubert et al. (2020); Arnold (2020); Lipsius (2018); Jarosch et al. (2019)

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

network theory to cluster workers and firms using survey data, however our microfoundations and detailed data allow us to identify more fine-grained heterogeneity and provide model-based interpretability of our classifications.

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

Roadmap: The paper proceeds as follows. Section 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 discusses our data and sample restrictions and presents basic summary statistics Section 5 demonstrates that our worker types and markets outperform traditional worker and job classifications in a variety of contexts. Section 6 applies our classifications to Bartik-style regressions and shows that standard methods may be understating the effects of shocks on workers. Section 7 concludes.

2 Model

In this section we develop a model based on Grigsby (2019) of workers supplying labor to jobs according to comparative advantage, where comparative advantage is determined by the interaction of potentially high-dimensional worker skills and job tasks. From this model of labor supply we derive our maximum likelihood estimator that clusters workers into worker types and jobs into markets. We embed the labor market model in a general equilibrium model with workers, jobs, firms, and a representative household comprised of workers that consumes firms' output. The general equilibrium model facilitates interpretation of the worker and job types and allows for counterfactuals in which we simulate labor demand shocks.

2.1 Labor supply

The labor market consists of workers, i, who supply labor to jobs, j. There is a unit mass of workers, each of whom belongs to one of I distinct worker types indexed by i. All workers in the same worker type are identical form the perspective of jobs. The exogenously determined

mass of type ι workers is denoted m_ι. Jobs are nested within firms and each job represents a set of tasks. A single job may employ multiple workers simultaneously. Each job belongs to one of Γ distinct markets indexed by γ and all jobs in the same market are identical from the perspective of workers.

For simplicity of exposition, we assume that worker types and markets are defined entirely by skills and tasks, respectively: all workers in the same worker type have the same set of skills and all jobs in the same market consist of the same set of tasks.[4] Type ι workers supply $\psi_{\iota\gamma}$ efficiency units of labor to jobs in market γ, where $\psi_{\iota\gamma}$ 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. See Grigsby (2019) for details.

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 are 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, which is determined in general equilibrium.[5] 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, 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

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

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*

[4] As we discuss in Section 3.3, it is straightforward to generalize worker types and markets to represent the intersection of skills/tasks, geography, preferences, credentials, and more.

[5] 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 (**?**).

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:

$$\mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu] = \frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)}. \tag{2}$$

We derive labor supply to market γ by aggregating equation (2) over all worker types, weighting by the mass of workers in worker type ι, m_ι, and the efficiency units of labor supplied by iota workers to gamma jobs, ψ_{ig}:

$$LS_\gamma(\vec{w}_t) = \sum_\iota m_\iota \mathbb{P}_\iota[\gamma_{it}|\Psi, \vec{w}_t, \Xi, \nu]\psi_{\iota\gamma} = \sum_\iota m_\iota \left(\frac{\exp\left(\frac{\psi_{\iota\gamma} w_{\gamma t} + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma' t} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma} \tag{3}$$

2.2 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 (1). We assume that the labor market parameters, $\{\Psi, \Xi, \nu\}$, and the demand shifters $\vec{a}$, are fixed across all T time periods we use for estimation (Assumption 2.3). 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.

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.*

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.

2.3 Firms

There are S sectors indexed by s. Each sector s consists of a continuum of firms producing the same sector-specific good 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, γ. The quantity of the sector s good produced by firm f, y_{sf}, is therefore given by

$$y_{sf} = \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} \tag{4}$$

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. 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 $\gamma(j)$. Therefore, the production function becomes

$$y_{sf} = \prod_{\{j\}_{j \in f}} \ell_j^{\beta_{\gamma s}}.$$

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 (4). Therefore, the firm solves

$$\pi_f = \max_{\{\ell_{\gamma f}\}_{\gamma=1}^{\Gamma}} \quad p_s \cdot \prod_{\gamma} \ell_{\gamma f}^{\beta_{\gamma s}} - \sum_{\gamma} w_\gamma \ell_{\gamma f}. \tag{5}$$

Production exhibits decreasing returns to scale because

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

where α denotes the labor share.

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 Household

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

$$U = \left(\sum_{s=1}^{S} a_s^{\frac{1}{\eta}} y_s^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}} \tag{6}$$

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

$$y_s^D = \frac{a_s}{\sum_{s'} \left(\frac{p_s}{p'_s} \right)^{\eta} (a_{s'} p_{s'})} Y \tag{7}$$

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

2.5 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. Our equilibrium has the following components:

1. The labor demand functions $\ell_{\gamma f}$ solve the firms' problem (5)

2. Labor supply is consistent with workers' expected utility maximization (2)

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: $y_s = \sum_{f \in s} \prod_{j \in f} \ell_j^{\beta_{\gamma s}}$ (4).

4. The labor market clears for each market γ: $LS_\gamma = LD_\gamma := \sum_s \sum_{f \in s} \sum_{j \in f | \gamma(j) = \gamma} \ell_j$

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

We solve the model numerically, as described in Appendix D.

3 Classifying workers and jobs

In this section, we start with the model of labor supply developed in the previous section and derive our maximum likelihood procedure for assigning workers to worker types, ι, and jobs to markets, γ. The only data used by the procedure is the set of realized worker–job matches. The procedure 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 (2), 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 (2), $\mathbb{P}_\iota[\gamma_{it} | \Psi, \vec{w_t}, \Xi, \nu]$, 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 (2) 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 market 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. Then job j's share of all market γ employment can be written

$$\mathbb{P}[j|\gamma] = \frac{d_j}{\sum_{j' \in \gamma} d^J_{j'}}. \tag{8}$$

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 (2) and equation (8):

$$\mathbb{P}_{ij} = \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)}}^{\mathbb{P}_i[\gamma|\Psi,\vec{w},\Xi,\nu]} \times \overbrace{\underbrace{\frac{1}{\sum_{j' \in \gamma} d^J_{j'}}}_{1/\text{ type } \gamma \text{ employment}} \times \underbrace{d_j}_{\text{Job } j \text{ employment}}}^{\mathbb{P}[j|\gamma]}. \tag{9}$$

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:

$$\begin{aligned}\mathbb{P}_{ij} &= \overbrace{\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)} \times \underbrace{\frac{1}{\sum_{j' \in \gamma} d^J_{j'}}}_{1/\text{ type } \gamma \text{ employment}}}^{:=\mathcal{P}_{\iota\gamma}} \times \underbrace{d_j}_{\text{Job } j \text{ employment}} \\ &= \mathcal{P}_{\iota\gamma} d_j.\end{aligned} \tag{10}$$

Define A_{ij} as 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, it follows that A_{ij} also follows a Poisson distribution:

$$A_{ij} \sim Poisson\left(d_i d_j \mathcal{P}_{\iota\gamma}\right). \tag{11}$$

For a complete proof, see Appendix H. Finally, define $\mathbf{A}$ as the matrix with typical element A_{ij}. $\mathbf{A}$ represents the full set of observed worker–job matches and is known as the adjacency matrix in network theory parlance. Since the elements of $\mathbf{A}$ are independent, we can write the density of $\mathbf{A}$ as

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d}_i,\vec{d}_j,\mathcal{P}\right)=\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{12}$$

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.

$$\vec{\iota},\vec{\gamma}=\underset{\substack{\{\vec{\iota}=\iota(i)\}_{i=1}^{N},\\ \{\vec{\gamma}=\gamma(j)\}_{j=1}^{J}}}{\arg\max}\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right) \tag{13}$$

This problem actually has five sets of parameters: the worker type and market assignments $\vec{\iota}$ and $\vec{\gamma}$, the worker and job match frequencies $\vec{d}_i$ and $\vec{d}_j$, and the type-specific match probabilities $\mathcal{P}_{\iota\gamma}$. 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. The worker and market assignments, $\vec{\iota}$ and $\vec{\gamma}$, are the parameters we choose in order to maximize the likelihood. Conditional on group assignments, the number of matches between each worker type–market pair is observable. Therefore, for a given set of $\vec{\iota}$ and $\vec{\gamma}$, we use these observed worker type-market match counts to compute observed match probabilities, which we use as our estimate of the true probabilities, $\mathcal{P}_{\iota\gamma}$. We iteratively update these parameters until we reach an optimum, as described below.

Equation (13) corresponds to the degree-corrected bipartite stochastic block model (BiSBM), a workhorse model in the community detection branch of network theory (see appendix B for details). It defines a combinatorial optimization problem, whereby 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 (13). This is not computationally feasible for large networks like ours. Therefore, we use a Markov chain Monte Carlo (MCMC) approach in which we modify the assignment of each worker to a worker type and each job to a market in a random fashion and accept or reject each modification with a probability given as a function of the change in the likelihood. We repeat the procedure for multiple different starting values to reduce the chances of finding local maxima. We implement the procedure using a Python package called graph-tool.

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

Equation (13) 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 penalizes a model that overfits the data by using a large number of parameters (corresponding to a large number of worker types and markets) by effectively adding a penalty term in our objective function, such that our algorithm finds a parsimonious model. This method has been found to work well in a number of real world networks (Peixoto, 2013; 2014b; Rosvall and Bergstrom, 2007). See appendix E for greater detail.

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 (13), 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.

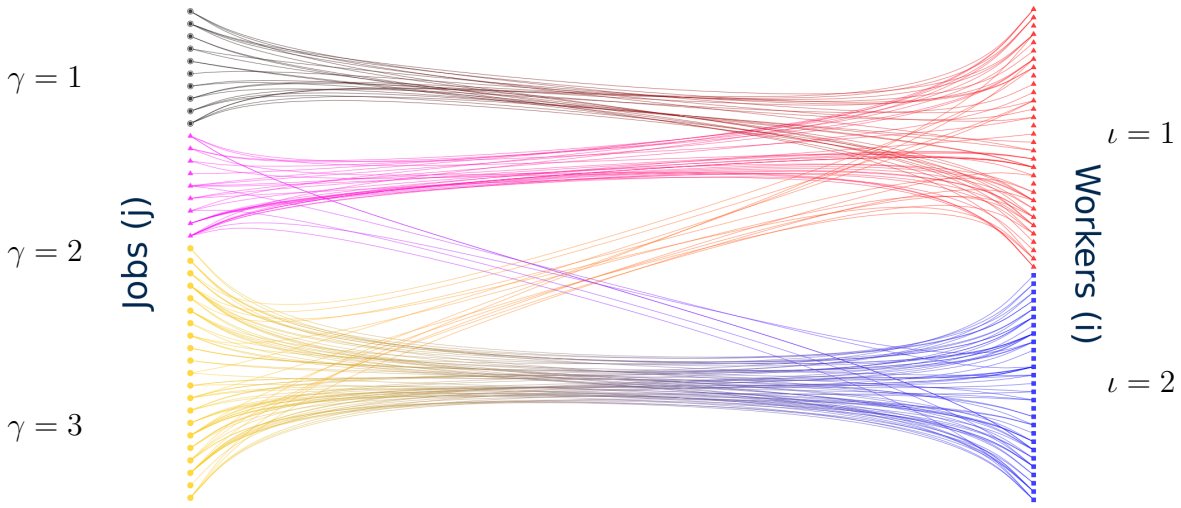
3.3 Discussion

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 2). 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 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 (12), which in turn implies a maximum likelihood estimation strategy. Now, we address the ramifications of these assumptions in turn.

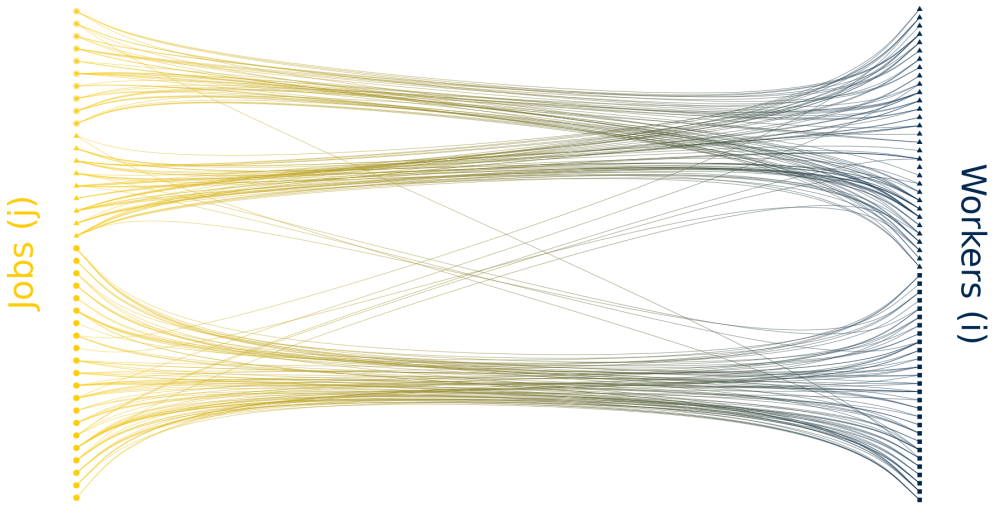
The first major assumption is that workers and jobs match according to a Roy model in which match probabilities are driven by skill-task match productivity. Since workers and jobs are clustered according to match probabilities, to the extent that match probabilities are determined by factors other than 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

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

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

discriminated against. Finally, 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 is key to identification because it implies that the set of worker–job matches is drawn independently 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.[6]

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.[7] An implication of the exogenous separations assumption is that a worker's match probabilities are independent of their job history, conditional on their type.[8]

[6]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.

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

[8]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

4 Data

We use the Brazilian linked employer-employee data set RAIS, which contains detailed data on all employment contracts in the Brazilian formal sector. Each observation in the data set represents a unique employment contract and includes a unique worker ID variable, an establishment ID, an occupation code, and earnings. Our sample includes all workers between the ages of 25 and 55 employed in the formal sector in the states of Sao Paulo, Minas Gerais, or Rio de Janeiro at least once between 2009 and 2012. These states are the 3 most populous and highest-GDP states in Brazil, are contiguous, and have a combined population of approximately 80 million; restricting to this set makes estimation computationally tractable. 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 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. Although we use occupation to define jobs, we do not use occupation as an input to our algorithm for classifying workers and jobs. This gives us a set of worker–job pairs 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. 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. We cannot dis-

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.

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

tinguish between non-employment and informal employment. Therefore, our outside option 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.[9] 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.[10]

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.

4.1 Summary statistics

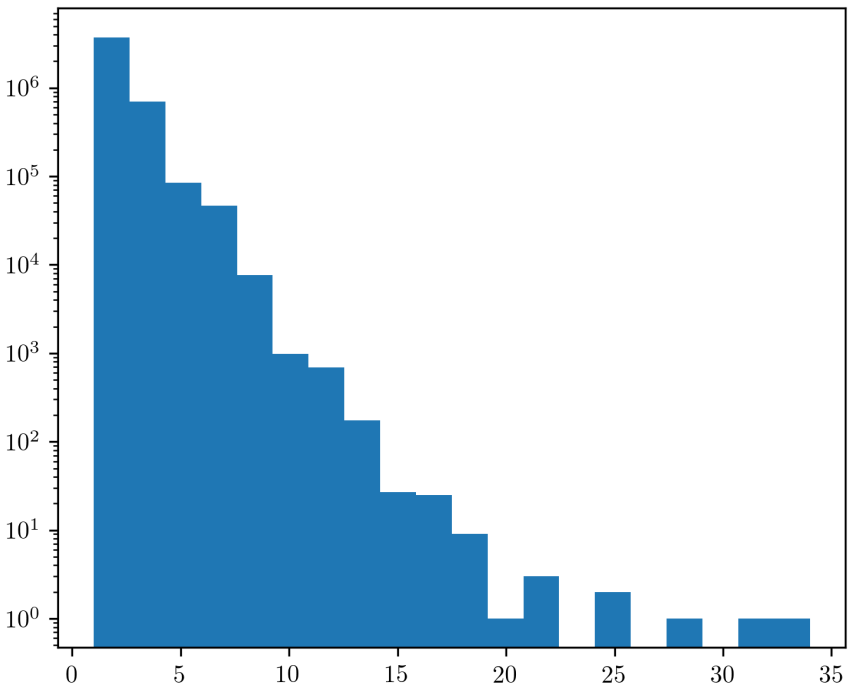
Our data contain 13,694,535 unique workers, 552,109 unique jobs, and 20,815,674 unique worker–job matches. The average worker matches with 1.73 jobs and the average job matches with 27.4 workers. 42% of workers match with more than one job during our sample. Figure 2 presents histograms of the number of matches for workers and jobs, respectively. In network theory parlance, these are known as degree distributions.

[9]IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Indicadores de subutilização da força de trabalho e de informalidade no mercado de trabalho brasileiro. Rio de Janeiro: IBGE, 2019.

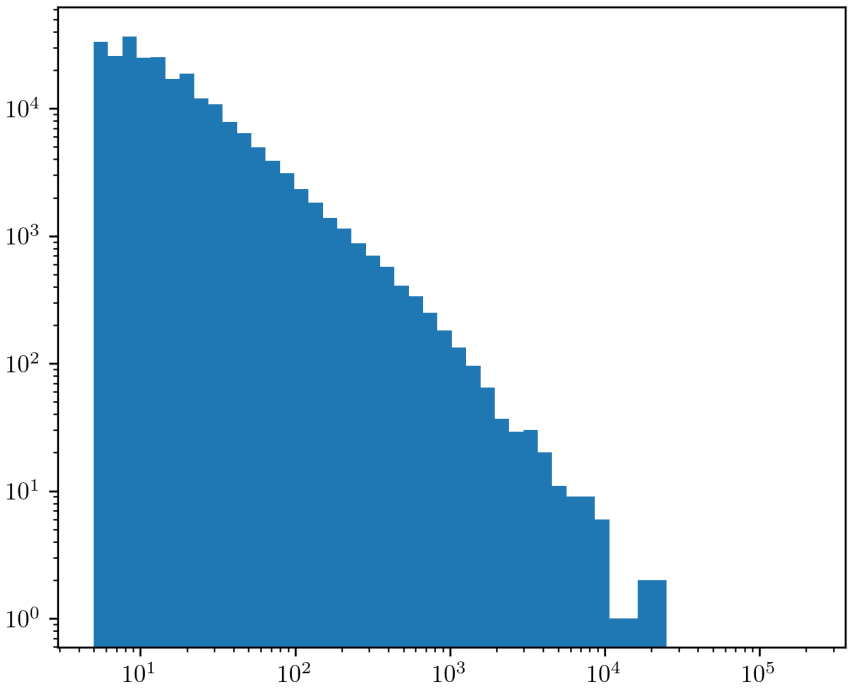
[10]See Engbom et al. (2021, Figure 21).

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 4.

Our network-based classification algorithm identifies 446 worker types (ι) and 1,371 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 unique workers and the median worker belongs to a worker type with 20,413 unique workers. The average job belongs to a market with 1,273 unique jobs and the median job belongs to a market with 1,188 unique jobs.

Table 2 presents the fraction of job changes that also involve a change in occupation, sector, market, firm, or establishment. The column "All Job Changes" computes the probability that a worker changes occupation, industry, sector, market, firm, or establishment conditional on changing jobs. The column "Firm Change Only" presents the same quantities restricting to the set of job changes that also involve a change in firm. The column "No Firm Change" restricts to job changes that do *not* involve a change in firm. Recall that we define a job as a 4-digit occupation–establishment pair. Table 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) demonstrates the value of allowing workers to costlessly change the market to which they supply labor, a feature that our model incorporates.

5 Validation of our worker types and markets

An ideal labor market definition is one that maximizes the similarity of workers and jobs within the same groups, respectively, and minimizes the similarity of workers and jobs in different groups. More formally, the ideal definition maximizes the within-market substitution elasticity and minimizes the cross-market substitution elasticity.[11] Directly estimating

[11]This is most easily understood if we extended our model in Section 2 to have a nested logit structure. Then, in addition to the parameter ν which governs the variance of idiosyncratic preference shocks *across* markets, we would have an additional parameter η that governs the correlation of idiosyncratic preference shocks *within* markets. To the extent that we incorrectly classify jobs incorrectly, ν and η will be biased towards each other. To understand this, consider the extreme case in which jobs are classified at random. Then the market classifications carry no economic information and, in expectation, two jobs in the same market will be equally substitutable from the worker's perspective as two jobs in different markets, meaning

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 4 using RAIS data from 2009–2012.

these elasticities is conceptually and computationally difficult, and is beyond the scope of this paper. Therefore, in this section we provide indirect evidence that our network-based classifications outperform standard classifications in identifying similar workers and similar jobs.

We provide five such pieces of evidence. First, we provide descriptive examples of the occupation and geographic distributions of jobs within our markets and demonstrate that they capture reasonable but potentially difficult to observe, forms of structure in the labor market that traditional market definitions miss. Second, we show that workers' labor supply is more concentrated within our markets than traditional definitions and jobs' hiring is more concentrated within our worker types. Third, we perform an out-of-sample prediction exercise in which our worker types and markets outperform traditional definitions in predicting workers' job-to-job flows. Fourth, we show that our worker types do a better job of identifying groups of workers with distinct skills as measured by the worker productivity parameter, ψ. Finally, we use the full general equilibrium extension of our model and show

that $\nu = \eta$. Alternatively, if we have perfectly classified the most similar jobs into the same markets, then jobs in the same market will be viewed as close substitutes (small η) and jobs in different markets will be more distant substitutes (large ν). We expand this argument and apply our market definitions to the question of measuring labor market power in ongoing work.

that our classifications more accurately predict the wage effects of the labor demand shock created by the 2016 Rio de Janeiro Olympics.

5.1 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[12] "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* different subsets of course instructors with other workers in different occupations who have similar skills. We pursue these examples in Tables 3 and 4.

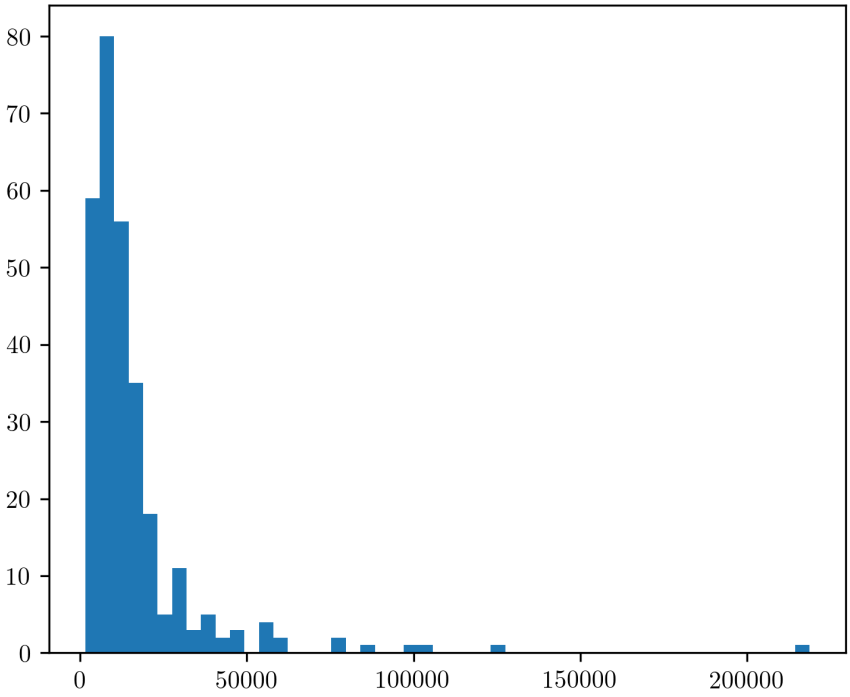
To illustrate this point more clearly, We focus on a specific worker type, $\iota = 17$, in which many workers are employed as course instructors. Table 3 presents the 10 occupations in which workers belonging to worker type $\iota = 17$ are most frequently employed.[13] The most frequently occurring- occupation is course instructr, however most of these occupations are related to physical fitness, education, or both. The fact that these course instructors tend to match with similar jobs as other workers whom we observe employed as personal trainers,

[12] Occupation names and descriptions are translated from Portuguese using Google Translate.

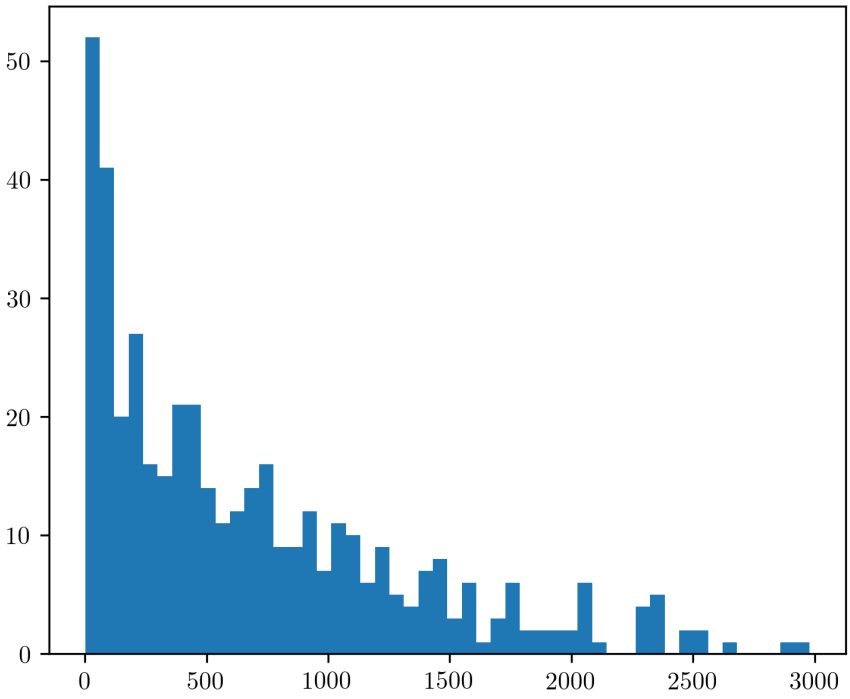
[13] To interpret this table, recall that we have assigned each individual worker to a worker type, ι. Each worker may be employed by one or more jobs in our sample, and each job is assigned an occupation code by the Brazilian statistical agency. A worker who has multiple jobs during the sample may have a different occupation associated with each job.

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.

physical education teachers, and sports coaches allows us to infer that these course instructors have skills more closely related to physical education than math.

Table 4 presents the distribution of occupations with a different worker type that contains many course instructors, $\iota = 52$. Unlike the previous example, where the workers appear to have physical education skills, the other frequently-occurring occupations in this worker type 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.

5.2 Worker types' labor market concentration

If our worker types and markets are successful in identifying groups of workers and jobs that are viewed as similar from the perspective of the labor market, then each worker type's labor supply will be concentrated within specific markets and each job's hiring will be concentrated among specific worker types. 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 classification schemes applied to the same underlying set of worker–job matches, the one that does a better job of identifying workers with similar skills and jobs requiring similar tasks will yield more concentrated labor supply and hiring distributions.

We compute each worker type's employment concentration across markets and occ2Xmesos

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 4 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 4 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.

using the Herfindahl-Hirschman index (HHI):

$$HHI_{\iota}^{Occ2Xmeso} = \sum_{s} \pi_{\iota s}^{2} \quad \text{and} \quad HHI_{\iota}^{Market} = \sum_{\gamma} \pi_{\iota \gamma}^{2}$$

where s indexes Occ2Xmesos, γ indexes markets, and $\pi_{\iota s}$ and $\pi_{\iota \gamma}$ are the share of type ι workers employed in occ2Xmeso 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 occ2Xmeso/market. Suppose we classified jobs randomly. Then worker types would not have a comparative advantage in specific markets and therefore would not be concentrated in specific markets. In this case, the HHI for each worker type would converge to $1/\Gamma$, where Γ is the total number of markets, 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 one market, 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 4a presents $HHI_{\iota}^{Occ2Xmeso}$ 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 Occ2Xmeso, which according to the argument above, indicates that markets identify groups of jobs that have more homogenous tasks than do sectors. Figure 4b repeats this analysis, instead focusing the hiring distribution of markets, classifying workers alternately by their worker type and by the first Occ2Xmeso in which they are ever ever observed. The story is analogous, with markets' hiring significantly more concentrated within worker types than within workers' first Occ2Xmeso.

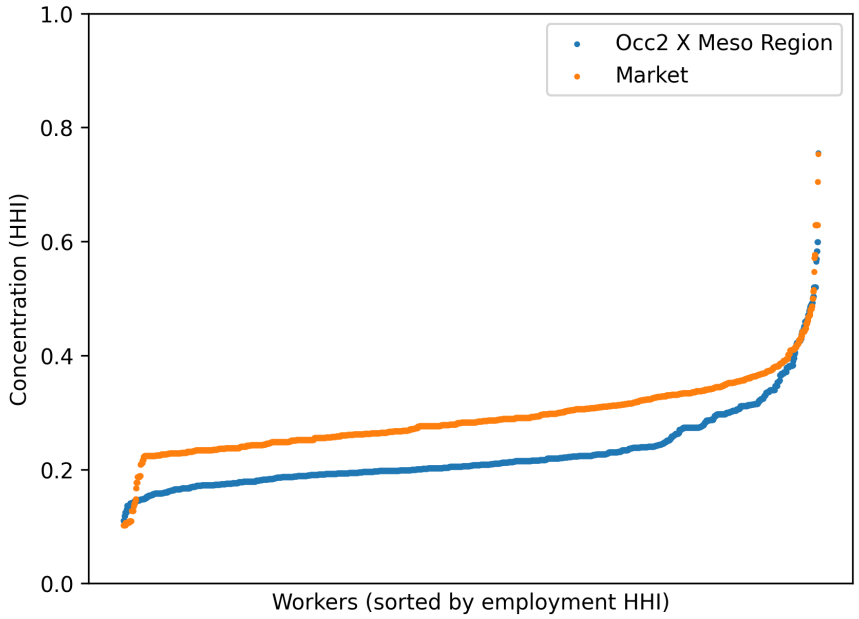
5.3 Predicting out-of-sample job-to-job flows

Another test of a market definition's success at identifying similar jobs is its ability to successfully predict job-to-job flows. As noted in Section 2, we assume that every time a worker changes jobs, they draw a new job from the same distribution. This implies that workers are more likely to transition to new jobs that have similar task distributions to their old jobs. Therefore, a market definition that does a better job of predicting job-to-job flows will be one that better captures latent job similarity. In this subsection we predict out-of-sample job-to-job flows using both our network-based market definitions and traditional definitions and show that our definitions outperform traditional ones.

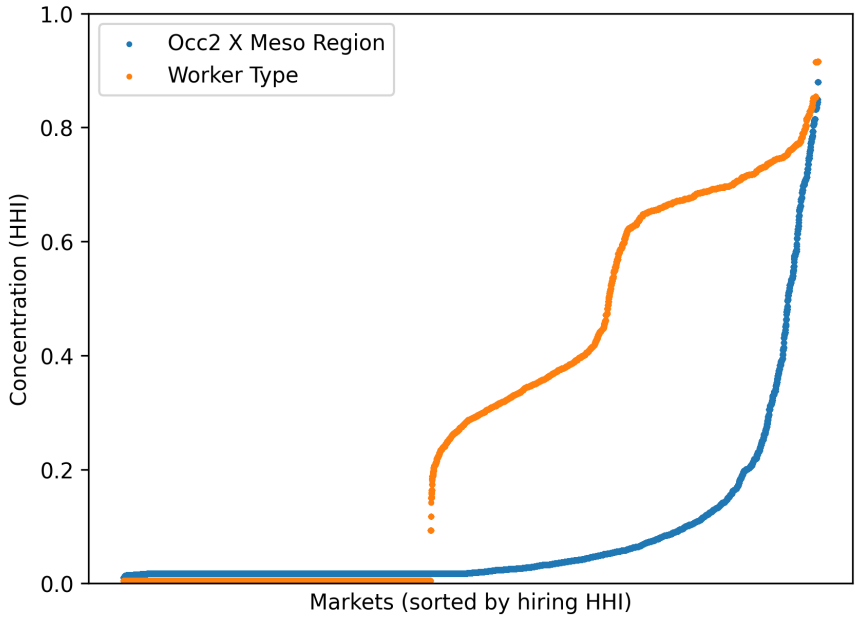
Consider a market definition $\mathbf{M}$ and the matrix of out-of-sample job-to-job empirical

Figure 4: Employment/Hiring Concentration According to Different Worker and Job Classifications

(a) Concentration of Worker Types' (ι) Employment Within Markets/Occ2Xmesos



(b) Concentration of Markets' (γ) Hiring Within Worker Types' (ι)/Occ2Xmesos



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 4.

transition probabilities $\mathbb{P}_{oos}[j'|j]$. The ability to predict future job-to-job transitions from the structure of $\mathbf{M}$ can be assessed as follows:

1. Compute a matrix of empirical market-to-market transition probabilities denoted by $\mathbb{P}_{\mathbf{M}}[m'|m]$

2. For each job j in market m, compute the probability of transitioning to j' in market m', $Prob(j' \in m'|j \in m)$, as follows:

$$Prob(j' \in m'|j \in m) := \mathbb{P}_{\mathbf{M}}[m'|m]\frac{d_{j'}}{\sum_{k \in m'} d_k}$$

   Intuitively, the probability that a worker in job j transitions to job j' is the product of the market-level transition probabilities $Prob(j' \in m'|j \in m)$ and the probability that job j' is chosen conditional on market m' being chosen. The latter probability is simply job j''s share of employment in market m', $\frac{d_{j'}}{\sum_{k \in m'} d_k}$.

3. Stack the probabilities computed in the previous step for each job, resulting in a predicted transitions matrix $\hat{\mathbb{P}}_{\mathbf{M}}[j'|j]$ with identical dimensions to the matrix $\mathbb{P}_{oos}[j'|j]$.

4. Compute a measure of fit, using a matrix norm of preference:

$$\Omega_{\mathbf{M}} := \left\| \hat{\mathbb{P}}_{\mathbf{M}}[j'|j] - \mathbb{P}_{oos}[j'|j] \right\|$$

We perform this exercise using empirical transitions from our estimation period of 2009–2011 and compute out-of-sample transitions using 2012 and 2013. We compare our network-based market definitions to a traditional alternative defined as the intersection of two-digit occupations and meso regions. We obtain lower prediction error using our network-based market definitions according to both the L_1 and L_2 norms. Specifically, the average L_1 prediction errors using markets and occ2Xmesos are 15.5 and 15.6, respectively. For the L_2 norm the corresponding values are 237.9 and 317.7. The greater discrepancy for the L_2 norm reflects the fact that predictions based on occ2Xmesos are much more likely to generate very large errors.

5.4 Worker type skill correlations

While Section 5.1 provided a qualitative example of our method's success in identifying clusters of workers with similar skills, we now provide quantitative evidence of our success in this regard. An ideal worker skills classification scheme will maximize the variance in skills

across different worker classifications and minimize the variance of skills within a worker classification. While we do not directly observe individual-level skills and therefore cannot measure within-classification skills variance, we do have a measure of across-classification skills variation. Each element of 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.

We estimate Ψ using maximum likelihood. Identification 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. For details, see Appendix C.

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, in expectation, 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 approximately 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. For this reason, we argue that if the pairwise correlations between different worker types' skills vectors tend to be close to zero, this is an indication that we have successfully identified groups of workers with distinct skills.

We summarize the correlations between different worker types' productivity vectors in Figure 5. We do this in two ways. In the left column we present correlation coefficients between all pairs of the $I = 446$ worker types in a lower triangular 446×446 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 above and detailed in Appendix C, except we classify workers and jobs by occupation and sector, rather than worker type and market. Row 2 of Figure 5 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 (446), allowing for comparisons at a similar level of granularity.

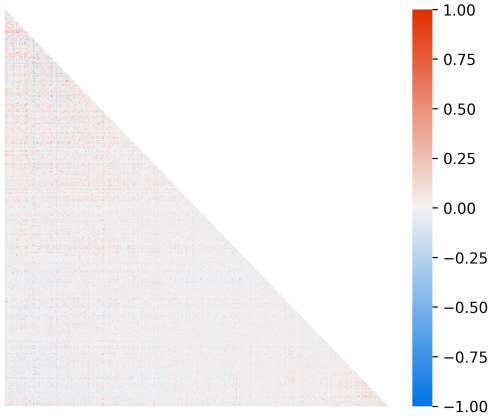
Figure 5 shows that correlations between different worker types' productivity vectors are smaller in magnitude when we use our model's (ι, γ) classifications rather than classifications based on labels available in the data, occupation and sector. This is because the network-based clusters of workers 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.

5.5 Predicting general equilibrium effects of Rio de Janeiro Olympics

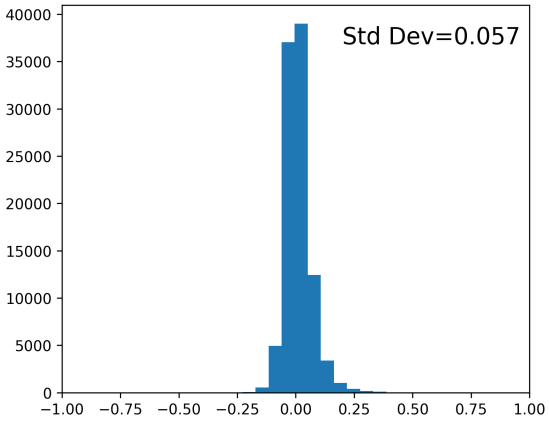
We test our model's ability to predict the effects of shocks in the context of the infrastructure investment and other preparations for the 2016 Rio de Janeiro Olympics. The Olympics were announced in late 2009 and construction of new venues and infrastructure were in full effect by 2014. Therefore,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

Figure 5: Skill Correlation Across Worker Types and Occupations

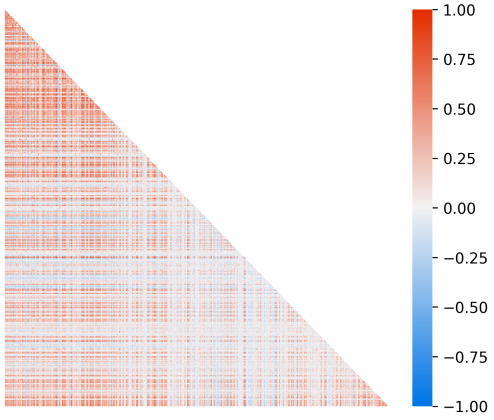
(a) (ι, γ) correlogram



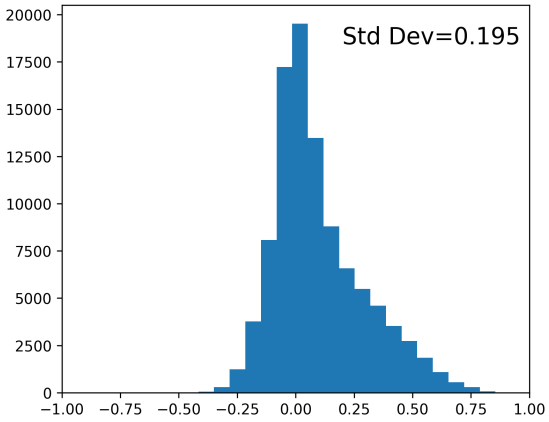
(b) (ι, γ) histogram



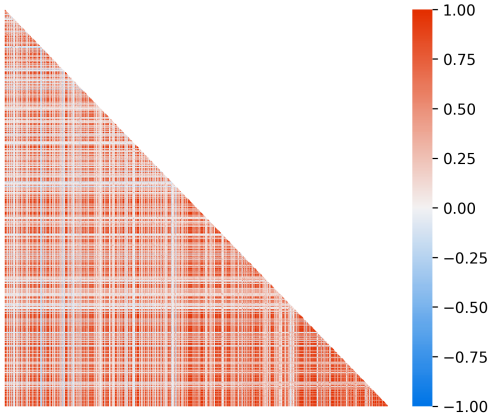
(c) (Occ4, Sector) correlogram



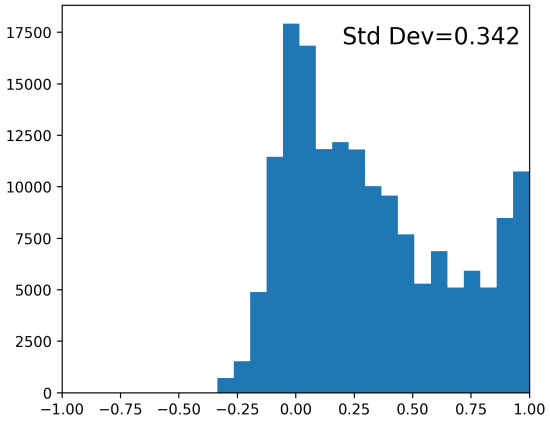
(d) (Occ4, Sector) histogram



(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 C.1 using the 2009-2012 RAIS worker earnings panel described in Section 4. 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).

on model-predicted changes in mean earnings for each worker type.

$$\Delta y = \beta_0 + \beta_1 \Delta \hat{y} + \varepsilon \tag{14}$$

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, respectively.[14]

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	-0.000	-0.002
	(0.009)	(0.009)	(0.009)	(0.011)	(0.010)
Model-implied Δ log earnings	0.982	0.148	0.428	0.234	0.560
	(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 (14) 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 (Maciente, 2013). 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 5.5. Dependent variables based on data from the 2009-2012 RAIS worker earnings panel described in Section 4. Independent computed by solving the model described in Section 2 using parameters estimated in Section C.1 and calibrated in Section C.2. Regressions are weighted by the number of workers per classification.

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

[14]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 4).

with fewer than 5,000 employees for computational reasons,[15] we are left with 306 4-digit occupations, yielding a level of disaggregation similar to the 446 worker types. The second two benchmarks characterize worker heterogeneity using k-means clusters of 6-digit occupations based on 225 O*NET skills, where the number of clusters is chosen to match the number of worker types ι, however we have to drop some of the resulting clusters because they are very small and are not observed in both the pre-shock and post-shock periods.[16] We classify job heterogeneity using sector in the first and third benchmark and using our network-based markets in the second and fourth. We present the results of these standard approaches in columns 2–5 of Table 5.

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

6 Reduced form estimation of labor market shocks

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

$$Bartik_g = \sum_s \pi_{gs} Shock_s \tag{15}$$

where g defines a group of workers, s defines a group of jobs, π_{gs} is the fraction of group g workers employed in group s jobs before the shock, and $Shock_s$ is the size of the shock to group s jobs. For example, in Autor et al.'s "China shock," g represents commuting zones, s indexes sectors, π_{gs} is commuting zone g's share of sector s employment, and $Shock_s$ is the growth in Chinese imports in sector s. $Shock_s$ is a proxy for the size of the labor demand shock in sector s jobs created by Chinese import growth, while π_{gs} governs which

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

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

workers are affected by the shock. Both $Shock_s$ and π_{gs} depend upon the researcher's choice of classifications, g and s, and therefore estimated effects of shocks are sensitive to these choices. In this section we study how the researcher's choice of worker and job classifications affect results.

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

6.1 Analysis of the 2016 Rio de Janeiro Olympics

We begin by once again considering the labor demand shock created by the preparations for the Rio de Janeiro Olympics. As in Section 5.5, we define 2009 as the pre-shock period and 2014 as the post-shock period. We regress 2009 to 2014 changes in worker group g earnings on the Bartik instrument defined in equation (15).

$$\Delta Earnings_g = \beta_0 + \beta_1 Bartik_g + \varepsilon_g \tag{16}$$

We have four specifications using all four combinations of our two worker classifications $g \in \{\text{worker type, occupation}\}$ and our two job classifications $s \in \{\text{market, sector}\}$. We normalize all of the Bartik instruments to have mean 0 and standard deviation 1 so that coefficients are directly comparable and can be interpreted as the effects of a 1 standard deviation change in the Bartik instrument on log earnings.[17] We measure π_{gs} as the fraction of group g workers who are employed in group s jobs. $Shock_s$ is alternatively defined as the change in sector-level product output or changes in the market-level labor input, ℓ_γ.

The results, presented in Table 6, show that estimated effects of the shock are highly sensitive to worker and job classifications. In column 1 we present our network-based classifications: workers are classified by worker type and jobs by market. In this specification, the effect of the shock on workers' earnings is positive and statistically significant, and the R^2 is large. The coefficient implies that a 1 standard deviation increase in exposure to the Olympics shock leads to an approximately 15.5% increase in earnings. Columns 2–4 present specifications using standard classifications. These specifications consistently find smaller

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

(and in some cases negative) effects of the shock on workers, and have less explanatory power for variation in worker earnings, as shown by the smaller R^2 values. These results are consistent with occupation and sector doing a worse job of characterizing worker skill and job task heterogeneity than worker types and markets, and this misclassification leading to attenuated estimates and worse model fit.

While our results indicate that classifying worker and job heterogeneity with error yields attenuated estimates of effects *in this case*, it is not necessarily the case that classification errors of this sort yield estimates that are biased towards zero in general. Since we do not have classical measurement error, the intuition of measurement error leading to attenuation bias does not apply. In fact, there is no theoretical prediction about the direction of the bias due to misclasssification of workers and jobs in our context (Mahajan, 2006; Hu, 2008). We confirm this through a series of simulations in which we generate a data set according to the data generating process implied by our model, randomly misclassify varying percentages of workers and jobs, and then estimate the Bartik regression, equation (16). We find no clear relationship between the amount of misclassification and the slope coefficient $\hat{\beta}$. However, we do find that the R^2 values decline approximately monotonically with the fraction of workers and jobs misclassified. Therefore, we interpret the larger R^2 values from estimating equation (16) using our network-based classifications as evidence that the network-based classifications classify worker and job heterogeneity with less error than the standard classifications. By contrast, the larger coefficient estimate when we use our network-based classifications is an empirical finding about the implications of misclassification in this context. See Appendix G for details on these simulations.

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

6.2 Reduced form analysis using simulated data

In this subsection, we demonstrate how estimated effects of shocks are sensitive to worker and job classifications in a setting where we can control the underlying data generating process. We replicate the analysis in the preceding section using simulated data. The simulated data have the same structure as the actual worker earnings panel described in Section 4 that we used to estimate the labor supply parameters and for the empirical exercises in Sections 5.5

Table 6: Effects of exposure to Rio Olympics shock

Exposure:	Market	Sector	Market	Sector	Market	Sector	Market	Sector	Market	Sector	Market
Worker classification:	ι	ι	ι	ι	ι	ι	Occ2×Meso Region	Occ2×Meso Region	Occ4	Occ4	ι
Job classification:	ι	N/A	ι	N/A	ι	N/A	Occ2×Meso Region	N/A	Occ4	N/A	γ
Intercept	-0.297***	-0.297***	-0.297***	-0.297***	-0.297***	-0.297***	-0.321***	-0.321***	-0.313***	-0.313***	-0.297***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.009)	(0.009)	(0.006)
Exposure (market)	0.058		0.027		0.027		0.017		0.038		0.058
	(0.006)		(0.006)		(0.006)		(0.007)		(0.009)		(0.006)
Exposure (sector)		-0.030		-0.030		-0.030		-0.024		-0.006	
		(0.006)		(0.006)		(0.006)		(0.007)		(0.009)	
Observations	446	446	446	446	446	446	1267	1267	570	570	446
R^2	0.198	0.051	0.041	0.051	0.043	0.051	0.004	0.008	0.030	0.001	0.198

Note:

Notes: Table presents the effect of the 2016 Rio de Janeiro Olympics shock on workers earnings from estimating equation (16). Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type (ι) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data from the 2009-2012 RAIS worker earnings panel described in Section 4 aggregated to the worker classification level.

and 6.1, and are drawn from the data generating process defined by our model. Since we control the data generating process, we can be certain that we are observing an exogenous labor demand shock that is unconfounded by, for example, concurrent labor supply changes.

We generate the simulated data as follows. First, we calibrate demand shifters $\vec{a}^{Pre}$ to match the levels of product demand in each sector in 2009. We then solve the model using the 2009 demand shifters to generate a pre-shock wage vector $\vec{w}^{Pre}$ that clears all markets γ. We draw worker types and four-digit occupations from the empirical joint distribution of worker types and four-digit occupations. To generate job matches for each worker recall that, conditional on searching, workers choose a market to supply labor to according to equation (1):

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

This implies that a type ι worker chooses market γ with probability given by equation (2):

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

where we use estimated parameter values $\hat{\Psi}$, $\hat{\Xi}$, and $\hat{\nu}$, estimated as described in Section C. All workers make this choice in period $t = 1$, and in subsequent periods workers search again if they draw a separation shock as described in Assumption 2.2. In our full model, workers match with individual jobs after choosing markets, however the identity of the worker's individual job j does not affect earnings or employment; it is only useful for classifying workers and jobs according to the BiSBM. Therefore, we do not specify the identity of each worker's specific job when generating our simulated data set.

Next, we draw sectors for each worker–job match according to the empirical joint distribution of sectors and markets. Finally, we draw earnings according to equation (22):

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

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

We repeat the four Bartik-style regressions from the previous section using our simulated data. The results, presented in Table 7, are qualitatively similar to the results using actual data in the previous section (Table 6), with the exception that the negative coefficients when

Table 7: Effects of exposure to *simulated* Rio Olympics shock

Exposure:	Market (γ)	Sector	Market (γ)	Sector
Worker classification:	Worker type (ι)	Worker type (ι)	Occ4	Occ4
Intercept	-0.239***	-0.239***	-0.264***	-0.264***
	(0.003)	(0.003)	(0.007)	(0.007)
iota exposure (market)	0.028			
	(0.003)			
iota exposure (sector)		0.022		
		(0.003)		
occ4 exposure (market)			0.004	
			(0.007)	
occ4 exposure (sector)				0.035
				(0.007)
Observations	446	446	573	573
R^2	0.158	0.099	0.001	0.045

Note:

Notes: Table presents the effect of the *simulated* 2016 Rio de Janeiro Olympics shock on workers earnings from estimating equation (16). Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type (ι) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data generated using our model as the data generating process, as described in Section 6.2, and aggregated to the worker classification level.

jobs are classified by sector are now small positive coefficients. We continue to find larger coefficients and R^2 values when we define shock exposure according to markets as opposed to sectors, and when we classify workers according to worker type as opposed to 4-digit occupation. These results reiterate our point that misclassifying worker and jobs causes us to significantly understate the effects of shocks on workers in this context. In the next section we demonstrate that this is a more general finding.

6.3 Simulating many shocks

In the previous sections we found that the estimated effects of shocks are larger when using our network-based worker and job classifications than when using standard classifications. To allay any concern that our finding is specific to the Rio Olympics shock, we replicate the analysis in the previous section for a series of different shocks. For each of the 15 sectors, we simulate a positive shock in which the demand shifter for the shocked sector is doubled and the demand shifters for all other sectors are unchanged, and a negative shock in which

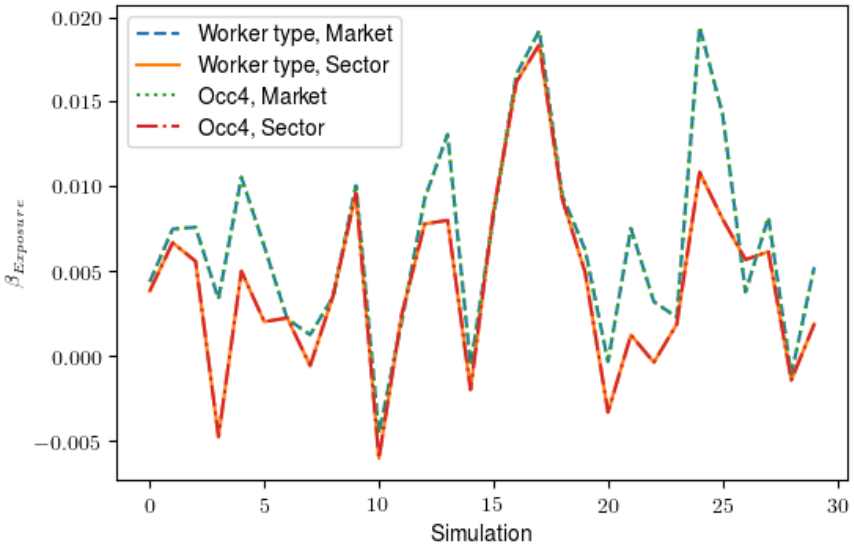
Table 8: Means across all simulated shocks

Worker Classification	Job Classification	Coefficient		R^2	
		Mean	Std Dev	Mean	Std Dev
Worker type	Market	0.007	0.006	0.043	0.052
Worker type	Sector	0.004	0.006	0.028	0.041
Occ4	Market	0.007	0.006	0.043	0.052
Occ4	Sector	0.004	0.006	0.028	0.041

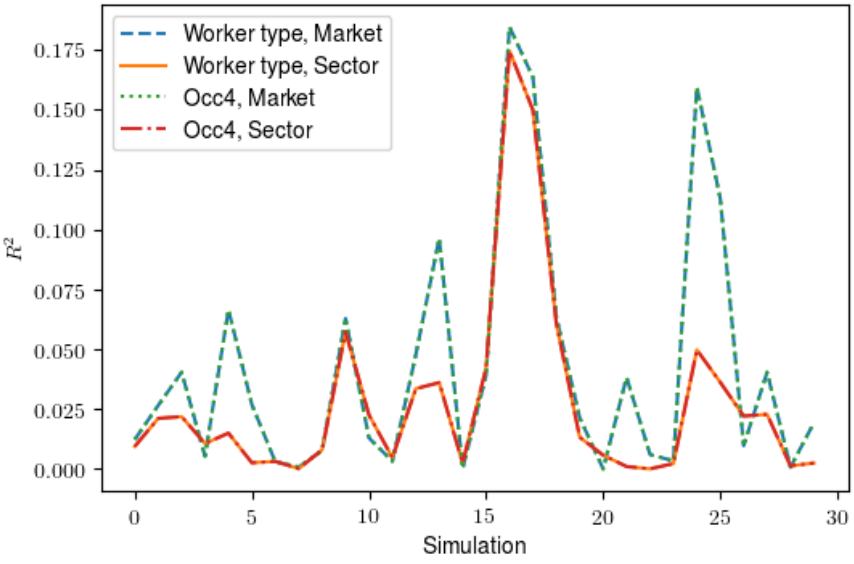
the demand shifter for the shocked sector is halved and the demand shifters for all other sectors are unchanged. For each shock, we generate a new simulated data set and then use the simulated data to estimate the Bartik-style regression in equation (16) for each of the four combinations of worker and job classifications: $g \in \{\text{worker type, occupation}\}$ and $s \in \{\text{market, sector}\}$. We present the results in Table 8. We consistently find larger coefficients and R^2 values using our network-based classifications. The average coefficient from our network-based classification specification is 3.7 times larger than the average coefficient from the occupation–sector specification, and the average R^2 is 11 times larger. Figure 6 presents the slope coefficients and R^2 values from each individual regression in these simulations and shows that our network-based classifications yield slope coefficients and R^2 values that are uniformly larger than those from standard classifications, not just larger on average.

Figure 6: Exposure coefficients from all simulated shocks

(a) Slope coefficients



(b) R^2 values



Notes: Figure presents estimated regression coefficients and R^2 values from estimating the Bartik-style regression, equation (16), for each of the 30 simulated shocks described in Section 6.3.

6.4 Case study of shock to the "Accommodations and Food" sector

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

Table 9 presents the same set of Bartik-style regressions as Tables 6 and 7 in the preceding sections. The qualitative story is unchanged: larger coefficients and R^2 values when we (i) define job heterogeneity according to markets as opposed to sectors, and (ii) when we define worker heterogeneity according to worker type as opposed to 4-digit occupation.

Table 9: Effects of exposure to simulated Accommodations and Food sector shock

Exposure:	Market (γ)	Sector	Market (γ)	Sector
Worker classification:	Worker type (ι)	Worker type (ι)	Occ4	Occ4
Intercept	-0.000	-0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.002)	(0.002)
iota exposure (market)	0.008			
	(0.001)			
iota exposure (sector)		0.006		
		(0.001)		
occ4 exposure (market)			-0.000	
			(0.002)	
occ4 exposure (sector)				-0.001
				(0.002)
Observations	290	290	306	306
R^2	0.092	0.054	0.000	0.001

Note:

Notes: Table presents the effect of the *simulated* Accommodations and Food sector shock on workers earnings from estimating equation (16). The shock is a 50% reduction in demand for the Accommodations and Food sector's output, holding demand for all other sectors' output constant. Independent variables normalized to have mean 0 and standard deviation 1. Workers classified by worker type (ι) in columns 1 and 2, and by 4-digit occupation in columns 3 and 4. Standard errors reported in parentheses. Jobs classified by market in columns 1 and 3, and by sector in columns 2 and 4. Estimated using data generated using our model as the data generating process, as described in Section 6.2, and aggregated to the worker classification level.

Table 10: Occupation counts for ι =64

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

Why does the Bartik instrument have more explanatory power for workers' outcomes when workers are classified by worker types and jobs are classified by markets? On the worker side, it is because, as we argued in Sections 5.1 and 5.4, our worker types do a better job of identifying groups of homogenous workers than do occupations. We see this again by focusing on one of the worker types that was most affected by the shock to the Accommodations and Food sector, worker type $\iota = 64$. Table 10 tabulates the 10 occupations we most frequently observe type $\iota = 64$ workers employed in. These occupations tend to be low-pay, low-education service sector occupations. The two most frequent are "food services assistant" and "retail salesperson." Our network-based classification method tells us that these retail and food services workers have similar skills despite the fact that they are employed in different occupations. If we had classified workers by occupation and jobs by sector, we would have implicitly assumed that the food services workers were exposed to the Accommodations and Food sector shock, while the retail salespeople were not. In reality, all of these workers were exposed to the shock because they have similar skills; workers not employed in the shocked sector may still be exposed to and affected by the shock if they are close substitutes for workers in the shocked sector. As we discussed in Section 6.1 and Appendix G, misclassifying workers such that some workers actually exposed to the shock are assumed not to have been exposed, and vice versa, leads to biased coefficient estimates and attenuated R^2 values.

On the jobs side, classifying jobs by market rather than sector more accurately captures the channels through which shocks propagate from jobs to workers. Bartik instruments based on standard classifications assume that workers supply labor directly to sectors; our classifications allow workers to supply labor directly to markets but only indirectly to sectors, by way of markets (see Figure 7). We illustrate why our approach is prefereable by again

focusing on type $\iota = 64$ workers. We have already established that these workers' skills are employable in both retail occupations and food service occupations, but who hires them? Do they supply labor to a retail market and a food services market? Or is there actually a market that includes jobs in both retail and food services? In Table 11 we present type $\iota = 64$ workers' labor supply by sector. Type $\iota = 64$ workers supply labor to a variety of sectors, including Retail, Wholesale and Vehicle Repair (28%) and Accommodations and Food (14%). Since these workers supply labor to such a variety of sectors, no single sector can reasonably approximate the set of jobs to which they supply labor. By contrast, type $\iota = 64$ workers' labor supply *is* concentrated within specific network-based markets, γ.

Table 12 presents the percentage of their labor that type $\iota = 64$ workers supply to each market, restricting to the top 10. Type $\iota = 64$ workers supply over 60% of their labor to a single market, market $\gamma = 47$, and there is no other market to which they supply more than 3.5 percent of their labor. In other words, type $\iota = 64$ workers' labor supply is highly concentrated within a specific market, but not nearly as concentrated in specific sectors, despite the fact that we have vastly more markets (1,371) than sectors (15). This is a specific example of the more general finding of greater concentration of employment within markets than sectors that we presented in Section 5.2. Worker types' employment is more concentrated within markets than sectors because our markets are designed to identify groups of jobs that compete for similar workers, whereas sectors are defined by product markets. Therefore, our markets more closely approximate the channels through which shocks propagate through the labor market to workers. By contrast, classifying jobs by sectors introduces error by grouping together jobs with heterogeneous changes in labor demand. Again, as we discussed in Section 6.1 and Appendix G, misclassifying jobs such that jobs that in fact hire dissimilar workers are assumed to hire similar workers, and vice versa, leads to biased coefficient estimates and attenuated R^2 values.

Table 11: Type $\iota = 64$ workers' labor supply by sector

Sector	Share (%)
Trade and repair of motor vehicles and motorcycles	28.5
Accommodation and food	12.7
Manufacturing industries	12.2
Professional, scientific and technical svcs	11.9
Private health and education	7.0
Arts, culture, sports and recreation and other ...	6.9
Transport, storage and mail	6.8
Construction	3.5
Utilities	3.1
Financial, insurance and related services	2.1
Extractive industries	1.8
Information and communication	1.8
Public admin, defense, educ, health and soc sec...	1.3
Real estate activities	0.2
Agriculture, livestock, forestry, fisheries and...	0.1

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

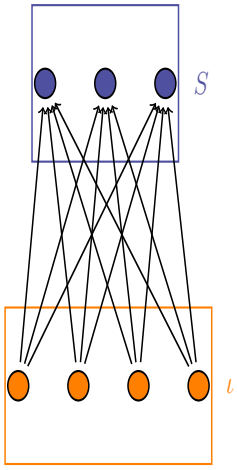
Table 12: Type $\iota = 64$ workers' labor supply by market (γ)

Market (γ)	Share (%)
47	60.1
189	3.5
116	1.7
242	1.5
418	1.3
83	1.3
36	1.2
138	1.1
125	0.9
45	0.8

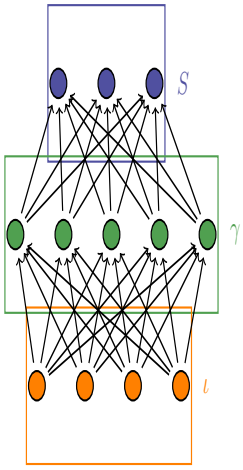
Notes: Table presents the share of type $\iota = 64$ workers employed in each market (γ) according to data generated by simulating the Accommodations and Food sector shock. The shock is a 50% reduction in demand for the Accommodations and Food sector's output, holding demand for all other sectors' output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

Figure 7

(a) Standard Classifications



(b) Our Model



7 Conclusion

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

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

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

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

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

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

References

Acemoglu, Daron and David Autor, "Skills, tasks and technologies: Implications for employment and earnings," 2011, *4*, 1043–1171.

Arnold, David, "Mergers and acquisitions, local labor market concentration, and worker outcomes," *Job Market Paper. https://scholar. princeton. edu/sites/default/files/d-harnold/files/jmp. pdf*, 2020.

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

_ , David Dorn, and Gordon H Hanson, "The China syndrome: Local labor market effects of import competition in the United States," *The American Economic Review*, 2013, *103* (6), 2121–2168.

_ , _ , _ , and Jae Song, "Trade adjustment: Worker-level evidence," *The Quarterly Journal of Economics*, 2014, *129* (4), 1799–1860.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Azar, Jose A., Ioana Marinescu, Marshall I. Steinbaum, and Bledi Taska, "Concentration in US Labor Markets: Evidence From Online Vacancy Data," Working Paper 24395, National Bureau of Economic Research March 2018.

Azar, Jose, Ioana Marinescu, and Marshall Steinbaum, "Measuring Labor Market Power Two Ways," 2019.

Bartik, Timothy J, "Who benefits from state and local economic development policies?," 1991.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim, "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?," Working Paper 24307, National Bureau of Economic Research February 2018.

Berger, David, Kyle Herkenhoff, and Simon Mongey, "Labor Market Power," *American Economic Review*, April 2022, *112* (4), 1147–93.

Blanchard, Olivier J. and Lawrence F. Katz, "Regional Evolutions," *Brookings Papers on Economic Activity*, 1992, *1*, 1–75.

Bonhomme, Stéphane, Kerstin Holzheu, Thibaut Lamadon, Elena Manresa, Thibaut Lamadon Elena Manresa, Magne Mogstad, and Bradley Setzler, "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?," 2020.

_ , Thibaut Lamadon, and Elena Manresa, "A distributional framework for matched employer employee data," *Econometrica*, 2019, *87* (3), 699–739.

_ , _ , and _ , "Discretizing Unobserved Heterogeneity," *arXiv preprint arXiv:2102.02124*, 2021.

Bound, John and Harry J Holzer, "Demand shifts, population adjustments, and labor market outcomes during the 1980s," *Journal of labor Economics*, 2000, *18* (1), 20–54.

Broda, Christian and David E Weinstein, "Globalization and the Gains from Variety," *The Quarterly journal of economics*, 2006, *121* (2), 541–585.

Burstein, Ariel, Eduardo Morales, and Jonathan Vogel, "Changes in between-group inequality: computers, occupations, and international trade," *American Economic Journal: Macroeconomics*, 2019, *11* (2), 348–400.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro, "Trade and labor market dynamics: General equilibrium analysis of the china trade shock," *Econometrica*, 2019, *87* (3), 741–835.

Card, David, "The impact of the Mariel boatlift on the Miami labor market," *ILR Review*, 1990, *43* (2), 245–257.

Engbom, Niklas, Gustavo Gonzaga, Christian Moser, and Roberta Olivieri, "Earnings Inequality and Dynamics in the Presence of Informality: The Case of Brazil," 2021.

Felix, Mayara, "Trade, labor market concentration, and wages," *Job Market Paper*, 2021.

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro et al., "Toward understanding the impact of artificial intelligence on labor," *Proceedings of the National Academy of Sciences*, 2019, *116* (14), 6531–6539.

Galle, Simon, Andres Rodriguez-Clare, and Moises Yi, "Slicing the pie: Quantifying the aggregate and distributional effects of trade," Technical Report, National Bureau of Economic Research 2017.

Gerlach, Martin, Tiago P Peixoto, and Eduardo G Altmann, "A network approach to topic models," *Science advances*, 2018, *4* (7), eaaq1360.

Grigsby, John, "Skill Heterogeneity and Aggregate Labor Market Dynamics," 2019.

Hu, Yingyao, "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 2008, *144* (1), 27–61.

Jarosch, Gregor, Jan Sebastian Nimczik, and Isaac Sorkin, "Granular search, market structure, and wages," Technical Report, National Bureau of Economic Research 2019.

Kantenga, Kory, "The effect of job-polarizing skill demands on the US wage structure," 2018.

Karrer, Brian and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E*, Jan 2011, *83*, 016107.

Kim, Ryan and Jonathan Vogel, "Trade shocks and labor market adjustment," *American Economic Review: Insights*, 2021, *3* (1), 115–30.

Larremore, Daniel B, Aaron Clauset, and Abigail Z Jacobs, "Efficiently inferring community structure in bipartite networks," *Physical Review E*, 2014, *90* (1), 012805.

Lindenlaub, Ilse, "Sorting multidimensional types: Theory and application," *The Review of Economic Studies*, 2017, *84* (2), 718–789.

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

Maciente, Aguinaldo, "The determinants of agglomeration in Brazil: input-output, labor and knowledge externalities." PhD dissertation, University of Illinois at Urbana-Champaign 2013.

Mahajan, Aprajit, "Identification and estimation of regression models with misclassification," *Econometrica*, 2006, *74* (3), 631–665.

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

Newman, MEJ, "Network structure from rich but noisy data," *Nature Physics*, 2018, p. 1.

Nimczik, Jan Sebastian, "Job Mobility Networks and Endogenous Labor Markets," 2018.

Peixoto, Tiago P, "Parsimonious module inference in large networks," *Physical review letters*, 2013, *110* (14), 148701.

_ , "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," *Physical Review E*, 2014, *89* (1), 012804.

Peixoto, Tiago P., "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks," *Phys. Rev. X*, Mar 2014, *4*, 011047.

Peixoto, Tiago P, "Nonparametric Bayesian inference of the microcanonical stochastic block model," *Physical Review E*, 2017, *95* (1), 012317.

_ , "Bayesian stochastic blockmodeling," *Advances in network clustering and blockmodeling*, 2019, pp. 289–332.

Rinz, Kevin, "Labor Market Concentration, Earnings Inequality, and Earnings Mobility," *CARRA Working Paper Series*, 2018, *2018* (10).

Rosvall, Martin and Carl T Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, 2007, *104* (18), 7327–7331.

Roy, Andrew Donald, "Some thoughts on the distribution of earnings," *Oxford economic papers*, 1951, *3* (2), 135–146.

Schmutte, Ian M., "Free to Move? A Network Analytic Approach for Learning the Limits to Job Mobility," *Labour Economics*, 2014, *29* (0), 49 – 61.

Schubert, Gregor, Anna Stansbury, and Bledi Taska, "Monopsony and outside options," *Available at SSRN*, 2020.

Sorkin, Isaac, "Ranking firms using revealed preference," *The quarterly journal of economics*, 2018, *133* (3), 1331–1393.

Tan, Joanne, "Multidimensional heterogeneity and matching in a frictional labor market - An application to polarization," 2018.

Yagan, Danny, "Employment Hysteresis from the Great Recession," Technical Report, National Bureau of Economic Research 2017.

Appendices

A Adding geography

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

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

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

B Network theory details

B.1 A primer on networks

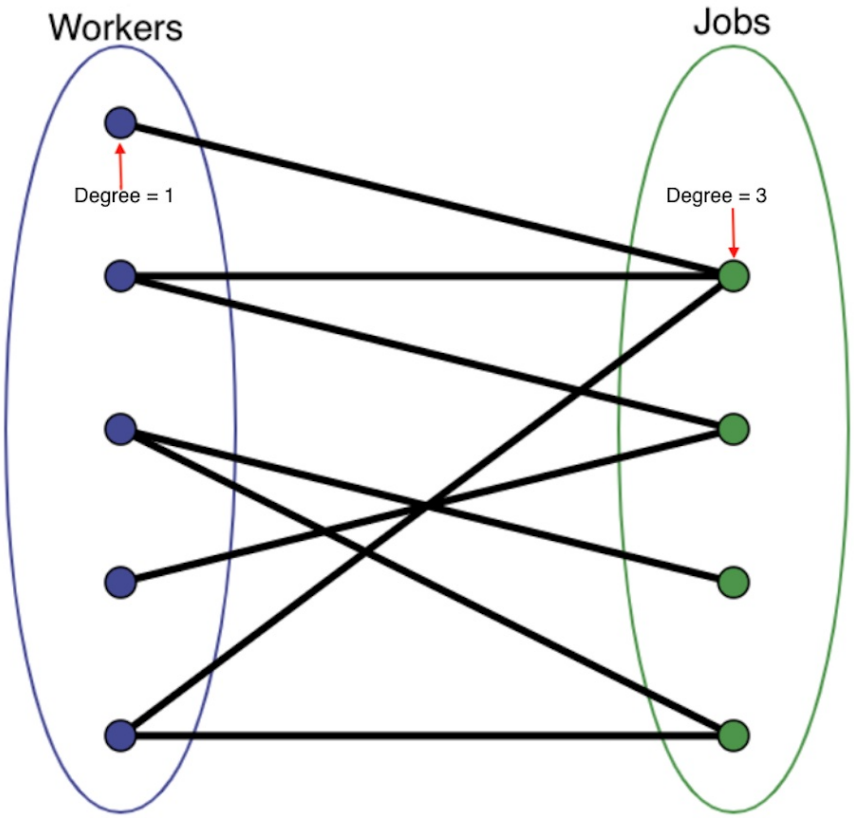
"A network is, in its simplest form, a collection of points joined together in pairs by lines" (Newman, 2018). The points are referred to as "nodes", and the lines as "edges." In Figure 8, the dots represent nodes and the lines represent edges. Networks can represent a wide variety of phenomena. For example, in an air travel network, airports are nodes and flight paths are edges. Similarly, in a social network, people are nodes and edges represent social relationships like friendship. The labor market, as viewed in LEED, can also be represented as a network. Each node represents an individual worker or job, and each edge represents an employment spell between a worker and a job.

In a network of worker–job connections like ours, edges connect workers to jobs. This means that there can be no edges between two worker nodes or between two job nodes; only between one worker node and one job node. Networks like this, in which nodes belong to one of two categories and all edges connect nodes in different categories, are known as "bipartite" networks. This is reflected in Figure 8 by the fact that all worker nodes are in blue on the left, all job nodes are in green on the right, and all edges (black lines) connect a worker to a job.

There is one more concept we need to introduce before returning our focus to estimation: the "degree" of a node. The degree of a node is the number of edges connected to that node. In figure 8, the first (from the top) worker node has a degree of 1 because it is connected to

exactly one edge (black line) while the first job node has a degree of 3. We index workers with i and jobs with j. We denote the degree of the node representing worker i d_i and the degree of the job representing job j d_j. In Figure 8, $d_{i=1} = 1$ and $d_{j=1} = 3$. As we discuss below, a worker who changes jobs more frequently will have a higher degree, while a job which hires more workers at a given time and/or has higher worker turnover will have a higher degree.

Figure 8: Simple bipartite network



In the next subsection, we show how our model generates a network of worker–job links similar to that in Figure 8, which can be observed using linked employer-employee data. Then, in the context of our model, we show how to back out latent worker types and markets from this observed network.

B.2 Bipartite Network Details

A network is a collection of nodes (also called "vertices"), connected to each other by edges. A *bipartite* network is a network in which there are two categories of nodes, and all edges connect a node of one category to a node of the other category. In our application, the two categories of nodes are workers and jobs, and all edges connect an individual worker to an

individual job. Alternatively, we could have defined a coworker network in which all of the nodes represent individual workers, and an edge connects pairs of workers who are coworkers. The coworker network is not a bipartite network because any node can be connected via an edge to any other node.

One way to represent a network is an adjacency matrix, typically denoted $\mathbf{A}$. The typical element of the adjacency matrix, A_{ij}, is the number of edges connecting nodes i and j. If there are n nodes in the network, then the adjacency matrix will have dimensions $n \times n$. In equation (17) below, we present an adjacency matrix for a bipartite network. Notice that there are two large blocks of zeros. This reflects the fact that edges only connect edges of different categories. In our case, edges only connect workers to jobs, not jobs to jobs or workers to workers. Suppose there are n_J jobs andd n_W workers, where $n_J + n_W = n$. Jobs are indexed by $(1, \ldots, n_J)$ and workers by $(n_J + 1, \ldots, n)$.

$$\mathbf{A} = \begin{array}{c} \overbrace{}^{\text{Jobs}} \quad \overbrace{}^{\text{Workers}} \\ \left(\begin{array}{ccccccc} 0 & \cdots & 0 & A_{1,n_J+1} & \ldots & A_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & A_{n_J,n_J+1} & \ldots & A_{n_J,n} \\ A_{n_J+1,1} & \ldots & A_{n_J+1,n_J} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & \ldots & A_{n,n_J} & 0 & \ldots & 0 \end{array}\right) \begin{array}{l} \left.\phantom{\begin{array}{c}x\\x\\x\end{array}}\right\} \text{Jobs} \\ \left.\phantom{\begin{array}{c}x\\x\\x\end{array}}\right\} \text{Workers} \end{array} \end{array}$$

We can also write the adjacency matrix as

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

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

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

B.3 Stochastic block model details

The *stochastic* in stochastic block model indicates that edges in the network are drawn stochastically from a data generating process (DGP). The *block* refers to the block structure of the DGP. Specifically, the SBM assumes that each node in the network belongs to a group $g \in 1, \ldots, G$. The probability of an edge between two nodes depends solely on group memberships of the two nodes.[18] Therefore, we can write a matrix of edge probabilities that has a block structure:

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

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

In this example, there are four nodes and two groups. Nodes 1 and 2 belong to group 1, as denoted by $g(1) = g(2) = 1$. Similarly, nodes 3 and 4 belong to group 2: $g(3) = g(4) = 2$. Instead of the edge probability matrix above, which can get quite large as the number of nodes grows, we can describe the matrix with two smaller objects: a vector indicating the group assignment of each node and a $G \times G$ matrix of group-specific edge propensities,[19] where G is the number of groups. We denote the vector of group assignments $\vec{g}$ and the matrix of group-specific edge propensities $\mathbf{\Omega}$. then

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

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

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

and

$$\mathbf{\Omega} = \begin{pmatrix} p_{g_1,g_1} & p_{g_1,g_2} \\ p_{g_2,g_1} & p_{g_2,g_2} \end{pmatrix} \tag{17}$$

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

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

The probability is slightly different for self-edges (edges connecting a node to itself):[20]

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

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

$$P(\mathbf{A}|\mathbf{\Omega}, \vec{g}) = \prod_{i<j} \frac{(\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\omega_{g_ig_j}\right) \times \prod_i \frac{(\frac{1}{2}\omega_{g_ig_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}\omega_{g_ig_i}\right) \tag{18}$$

Although equation (18) presents the standard SBM, this formulation is rarely used in practice. For empirical applications, researchers typically use an extension called the *degree-corrected* stochastic block model (DCSBM). The difference between the SBM and the DCSBM is that the DCSBM allows the expected degree of each node (the number of edges connected to that node) to vary. This more-closely matches real world data and the DCSBM has been shown to have far superior performance in empirical applications than the SBM (Karrer and Newman, 2011). Let $\vec{d}$ be vector containing the degree of each node, with typical element d_i representing the degree of node i. We can write the DCSBM as

$$P(\mathbf{A}|\vec{d}, \mathbf{\Omega}, \vec{g}) = \prod_{i<j} \frac{(d_id_j\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_id_j\omega_{g_ig_j}\right) \times \prod_i \frac{(\frac{1}{2}d_i^2\omega_{g_ig_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}d_i^2\omega_{g_ig_i}\right). \tag{19}$$

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

B.4 Community detection using the stochastic block model

In Section B.3 we assumed that we know all of the parameters of the model: $\vec{d}$, $\mathbf{\Omega}$, and $\vec{g}$. However, in actual applications, we typically observe the network $\mathbf{A}$ and the degree distribution $\vec{d}$ and want to recover the group memberships of the nodes $\vec{g}$. (Conditional on knowing $\vec{g}$, we can also compute the empirical edge probabilities matrix $\hat{\mathbf{\Omega}}$.) Therefore, we recover the group memberships of the nodes, $\vec{g}$, by treating equation (19) as a maximum likelihood problem and choosing the group memberships in order to maximize the probability of the observed adjacency matrix $\mathbf{A}$, given the data. We write the likelihood

$$\mathcal{L}(A|\vec{g}) = \prod_{i<j} \frac{(d_i d_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp\left(-d_i d_j \omega_{g_i g_j}\right) \times \prod_i \frac{(\frac{1}{2} d_i^2 \omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2} d_i^2 \omega_{g_i g_i}\right) \quad (20)$$

and our task is to choose

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

B.5 Bipartite stochastic block model details

The bipartite stochastic block model (BiSBM) is an extension of the SBM (Section B.3) applied to bipartite networks (Section B.2). The edge probability matrix has the same block structure as in the SBM, however since it is a bipartite network, there are two categories of nodes — in our case workers and jobs — and all edges connect a node from one category (a worker) to a node from the other (job).

Suppose there are two types of workers, indexed by $\iota \in 1, 2$, and two types of jobs, indexed by $\gamma \in 1, 2$. Suppose further that there are 4 individual workers and 4 individual jobs, indexed by $i = 1, \ldots, 4$ and $j = 1, \ldots, 4$, respectively. There are two individual workers and two individual jobs of each type. Denote the probability of an edge between a type ι

worker and a job in market γ as $\omega_{\iota\gamma}$. Then we have the following edge probability matrix

$$
\begin{array}{c}
\\
\\
\\
j=1,\gamma=1\\
j=2,\gamma=1\\
j=3,\gamma=2\\
j=4,\gamma=2\\
i=1,\iota=1\\
i=2,\iota=1\\
i=3,\iota=2\\
i=4,\iota=2
\end{array}
\begin{array}{c}
\overbrace{\qquad\qquad\qquad\qquad}^{\text{Jobs}}\quad\overbrace{\qquad\qquad\qquad\qquad}^{\text{Workers}}\\
\begin{array}{cccccccc}
j=1 & j=2 & j=3 & j=4 & i=1 & i=2 & i=3 & i=4\\
\gamma=1 & \gamma=1 & \gamma=2 & \gamma=2 & \iota=1 & \iota=1 & \iota=2 & \iota=2
\end{array}\\
\begin{pmatrix}
0 & 0 & 0 & 0 & \omega_{11} & \omega_{11} & \omega_{21} & \omega_{21}\\
0 & 0 & 0 & 0 & \omega_{11} & \omega_{11} & \omega_{21} & \omega_{21}\\
0 & 0 & 0 & 0 & \omega_{12} & \omega_{12} & \omega_{22} & \omega_{22}\\
0 & 0 & 0 & 0 & \omega_{12} & \omega_{12} & \omega_{22} & \omega_{22}\\
\omega_{11} & \omega_{11} & \omega_{12} & \omega_{12} & 0 & 0 & 0 & 0\\
\omega_{11} & \omega_{11} & \omega_{12} & \omega_{12} & 0 & 0 & 0 & 0\\
\omega_{21} & \omega_{21} & \omega_{22} & \omega_{22} & 0 & 0 & 0 & 0\\
\omega_{21} & \omega_{21} & \omega_{22} & \omega_{22} & 0 & 0 & 0 & 0
\end{pmatrix}
\end{array}
\begin{array}{l}
\\
\\
\\
\\
\\
\\
\\
\\
\left.\begin{array}{l}\\\\\\\\\end{array}\right\}\text{Jobs}\\
\left.\begin{array}{l}\\\\\\\\\end{array}\right\}\text{Workers}
\end{array}
\begin{array}{l}
\\
\\
\}\text{Worker/Job Index}\\
\}\text{Worker/market}
\end{array}
.
$$

The primary takeaway from this matrix is that the probability of a connection between a pair of nodes is determined by their group memberships. If worker i belongs to type ι and job j belongs to type γ, then the probability of worker i matching with job j is governed by $\omega_{\iota\gamma}$. The two blocks of zeros in this matrix reflect the fact that the probability of an edge between two workers or two jobs is zero in a bipartite network.

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

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

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

$$P(\mathbf{A}|\vec{d}, \mathbf{\Omega}, \vec{g}) = \prod_{i<j}\frac{(d_i d_j \omega_{g_ig_j})^{A_{ij}}}{A_{ij}!}\exp\left(-d_i d_j \omega_{g_ig_j}\right). \tag{21}$$

Notice that this expression lacks the second term found in equation (19), which captures self-edges in which an edge runs connects a node to itself. This is because self-edges are impossible in a bipartite network, since self-edges would connect nodes belonging to the same category (e.g. workers to workers).

B.6 Visual representation of linked employer-employee data as a network

Our raw data looks like what is presented in Table 13, with the exception that we generate the "JobID" column ourselves by concatenating the establishment code ('Estab Code') and occupation code ('Occ Code'). However, we only use the two variables 'WorkerID' and 'JobID' in estimation. Therefore, in Figure 9, we show the worker and job IDs from the data alongside a network representation of the same data. In the network representation, workers are blue dots on the right, jobs are yellow dots on the left, and black lines represent edges connecting workers to jobs at which they were employed. Finally, in Table 14, we present an adjacency matrix representation of the same network.

Table 13: Sample linked-employer-employee data

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

Figure 9: Representing the data as a network

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

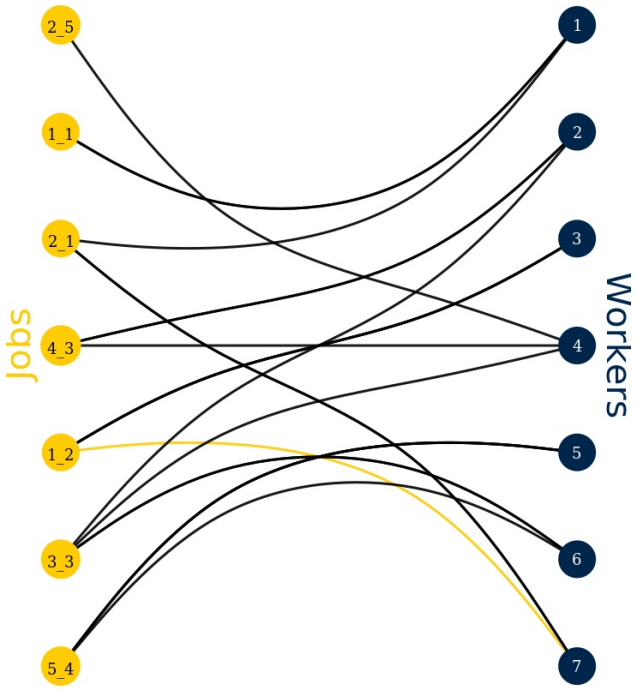


Table 14: Adjacency matrix: **A**

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

C 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 Section 3.

C.1 Estimating Ψ from observed matches

Identification and estimation of the labor supply parameters builds upon Bonhomme et al. (2019) and Grigsby (2019), with the key difference being that we assign both workers to worker types and jobs to markets prior to estimating labor supply parameters and do so in a way that more fully exploits the information revealed by worker–job matches, allowing us to identify a significantly greater degree of worker and job heterogeneity.[21]

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

$$\omega_{it} = \psi_{\iota(i)\gamma_{it}} w_{\gamma_{it}} e_{it}. \tag{22}$$

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

Assumption C.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 C.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.

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

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 worker i choosing a job in market γ and comes from the Roy model (equation 2). 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}.[22] Assumptions 2.2 and C.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:

$$\begin{aligned}
\mathcal{L}(\Theta|\mathbb{X}) &= \prod_{i=1}^{N}\prod_{t=1}^{T} l(\gamma_{it}, \omega_{it}, c_{it}|\mathbb{X}) \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T} \underbrace{[\mathbb{P}(\gamma_{it}|\Theta) f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{Separation}}^{c_{it}} \underbrace{[f_\omega(\omega_{it}|\iota(i), \gamma_{it}, \Theta)]}_{\text{No separation}}^{1-c_{it}} \qquad (23)
\end{aligned}$$

[22] 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.

Finally, the log-likelihood is

$$\ell(\Theta|\mathbb{X}) = \sum_{i=1}^{N}\sum_{t=1}^{T} c_{it} \log \mathbb{P}(\gamma_{it}|\Theta) + \sum_{i=1}^{N}\sum_{t=1}^{T} \log f_{\omega}(\omega_{it}|\iota(i), \gamma_{it}, \Theta) \tag{24}$$

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

Assumption C.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 C.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 C.2 assumes that wages follow a log-normal distribution which is worker type-market specific, following Bonhomme et al. (2019) and Grigsby (2019). Assumption C.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 C.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.[23]

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

[23]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.

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 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 F.

C.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.[24]
- 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).

[24]Broda and Weinstein (2006) estimate this parameter to be 4, however their estimate comes from significantly more disaggregated product categories, so we choose a smaller value. This parameter affects our structural results in Section 5.5, but does not affect the reduced form estimates in Section 6.

D Model Solution Appendix

Firm's problem

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

$$\max_{\ell_{\gamma s}} \quad p_s \prod_{\gamma} \ell_{\gamma s}^{\beta_{\gamma s}} - \sum_{\gamma} w_\gamma \ell_{\gamma s} \tag{25}$$

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

$$\ell_{\gamma s}^D = \frac{p_s \beta_{\gamma s} \left(\prod_{\gamma'} \ell_{\gamma' s}^{D\ \beta_{\gamma' s}} \right)}{w_\gamma} \tag{26}$$

Combining the Γ FOCs for a given sector S:

$$\ell_{\gamma s}^D = \frac{\beta_{\gamma s}}{\beta_{\gamma' s}} \frac{w_{\gamma'}}{w_\gamma} \ell_{\gamma' s}^D \tag{27}$$

Plugging in 27 for $\ell_{\gamma s}^D$ in equation 26, we have

$$\ell_{\gamma s}^D = \left[p_s \left(\frac{\beta_{\gamma s}}{w_\gamma} \right)^{1-\sum_{\gamma'} \beta_{\gamma' s}} \prod_{\gamma'} \left(\frac{\beta_{\gamma' s}}{w_{\gamma'}} \right)^{\beta_{\gamma' s}} \right]^{\frac{1}{1-\sum_{\gamma'} \beta_{\gamma' s}}} = \ell_{\gamma s}^D \left(\vec{p}, \vec{w} \right) \tag{28}$$

which represents labor demand for firm s, using only FOCs for firm s.[25]

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

$$y_s^S = y_s^S \left(\{ \ell_{\gamma s}^D (\vec{p}, \vec{w}) \}_{\gamma=1}^{\Gamma} \right) = \prod_{\gamma} \ell_{\gamma s}^{D\ \beta_{\gamma s}} \tag{29}$$

Household's problem

[25] We could alternatively write this expression as

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

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

Lagrangean:

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

FOC:

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

Simplifying,

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

Rearranging,

$$y_s^D = \frac{U}{\lambda^\eta} \frac{a_s}{p_s^\eta} \tag{30}$$

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

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

Plugging this into 30, we have our expression for product demand:

$$y_s^D = \frac{a_s Y}{p_s^\eta \sum_{s'} \left(a_{s'} p_{s'}^{1-\eta}\right)} = y_s^D(\vec{p}, Y) \tag{31}$$

Worker's problem

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

Solving the worker's problem gives labor supply:

$$\ell_\gamma^S(\vec{w}) = \sum_\iota m_\iota \left(\frac{\exp\left(\frac{\psi_{\iota\gamma} w_\gamma + \xi_\gamma}{\nu}\right)}{\sum_{\gamma'=0}^{\Gamma} \exp\left(\frac{\psi_{\iota\gamma'} w_{\gamma'} + \xi_{\gamma'}}{\nu}\right)} \right) \psi_{\iota\gamma} \tag{32}$$

Equilibrium

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

1. Labor market:

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

2. Product market:

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

3. Spending = Income = Wages + Profits

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

where

1. Product demand:

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

2. Product supply:

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

3. Labor supply:

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

4. Labor demand:

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

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

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

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

1. Choose vectors of start values for wages $\vec{w}$ and prices $\vec{p}$
2. Compute labor supply $\ell_\gamma^S(\vec{w})$ given wages $\vec{w}$ following equation 32
3. Compute labor demand $\ell_{\gamma s}^{D}(\vec{p}, \vec{w})$ given these start values following equation 28
4. Compute the product supply $y_s^S\left(\{\ell_{\gamma s}^{D}(\vec{p}, \vec{w})\}_{\gamma=1}^{\Gamma}\right)$ implied by the labor demand choice in the previous step following equation 29
5. Compute household income $Y = \sum_s p_s y_{\gamma s}^{S}$ implied by product supply in the previous step
6. Compute product demand $y_s^D(\vec{p}, Y)$ following equation 31
7. Update prices using the update rule $p_s^{t+1} = p_s^t \left(\frac{y_s^D}{y_s^S} \right)^\rho$, where ρ is a dampening factor that controls the size of the update and t indexes iterations. Intuitively, we increase prices if demand exceeds supply, and decrease them if supply exceeds demand. The size of the update depends on the size of the mismatch between supply and demand.
8. Update wages using the update rule $w_\gamma^{t+1} = w_\gamma^t \left(\frac{\ell_\gamma^D}{\ell_\gamma^S} \right)^\rho$
9. Repeat steps 2-8 until convergence

E Choosing number of worker types and markets

Equation 12 defined the probability of observing our network of worker–job matches, denoted by the adjacency matrix $\mathbf{A}$:

$$P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right)=\prod_{i,j}\frac{\left(d_i d_j \mathcal{P}_{\iota(i)\gamma(j)}\right)^{A_{ij}}}{A_{ij}!}\exp\left(d_i d_i^J \mathcal{P}_{\iota(i)\gamma(j)}\right). \tag{33}$$

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

$$P\left(\mathbf{A},\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right)=P\left(\mathbf{A}\middle|\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right)P\left(\vec{d_i},\vec{d_j}\middle|\vec{\iota},\vec{\gamma},\mathcal{P}\right)P\left(\mathcal{P}\middle|\vec{\iota},\vec{\gamma}\right)P\left(\vec{\iota},\vec{\gamma}\right) \tag{34}$$

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

It turns out that this Bayesian formulation has an equivalent information-theoretic interpretation. We can rewrite the joint probability defined in equation (34) as

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

where

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

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

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

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

$$\mathcal{L}=-\log_2 P\left(\vec{\iota},\vec{\gamma},\vec{d_i},\vec{d_j},\mathcal{P}\right)$$

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

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

We choose the assignment of workers to worker types and jobs to markets that maximizes the posterior of the distribution, equation (34). This is equivalent to choosing the set of parameters that yields the smallest description length, and therefore compresses the data the most. Intuitively, we can think of $\mathcal{L}$ as a penalty term that increases with the number of parameters, and thereby prevents overly complex models. If the number of worker types and markets becomes large, $\mathcal{S}$ will increase, indicating a better model fit, but the penalty term $\mathcal{L}$ will increase as well. The chosen model will therefore be the one that maximizes the quality of the model fit relative to the cost imposed by the penalty term.

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

F Identification of Labor Supply Parameters

Taking the first order conditions of equation 24 with respect to each of the parameters provides intuition for how the parameters are identified.

F.1 ν

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

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

will be large.

F.2 ξ_γ

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

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

F.3 $\phi_{\iota\gamma}$

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

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

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

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

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

F.4 λ

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

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

F.5 σ

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

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

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

G Measurement error

The Bartik regressions in equation 16 can be written

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

where

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

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

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

$$Y = X\beta + U$$

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

$$\begin{aligned}\tilde{X} &= X + V_X \\ \tilde{Y} &= Y + V_Y\end{aligned}$$

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

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

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

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

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

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

The true coefficient β can be written

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

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

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

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

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

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

Therefore,

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

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

We demonstrate this point through a simulation. We simulate a shock as described in Section 6.2. We estimate a series of regressions on changes in earnings by worker type on the Bartik instrument with jobs classified by market, however in each regression we randomly misclassify some percentage of workers and jobs. We loop from 0 to 100 percent of workers misclassified in intervals of five percent, and within each loop perform the same loop from 0 to 100 percent of jobs misclassified. We present the coefficients on the Bartik instrument in Figure 10 and the R^2 values in Figure 11. R^2 values decline approximately monotonically with the degree of misclassification in both the worker and job dimensions, as expected. By contrast, there is much less of a coherent story with the regression coefficients. Again, this is consistent with the theoretical prediction that the effect of misclassification on regression coefficients is indeterminate.

Figure 10: Coefficient estimates with worker and job misclassification

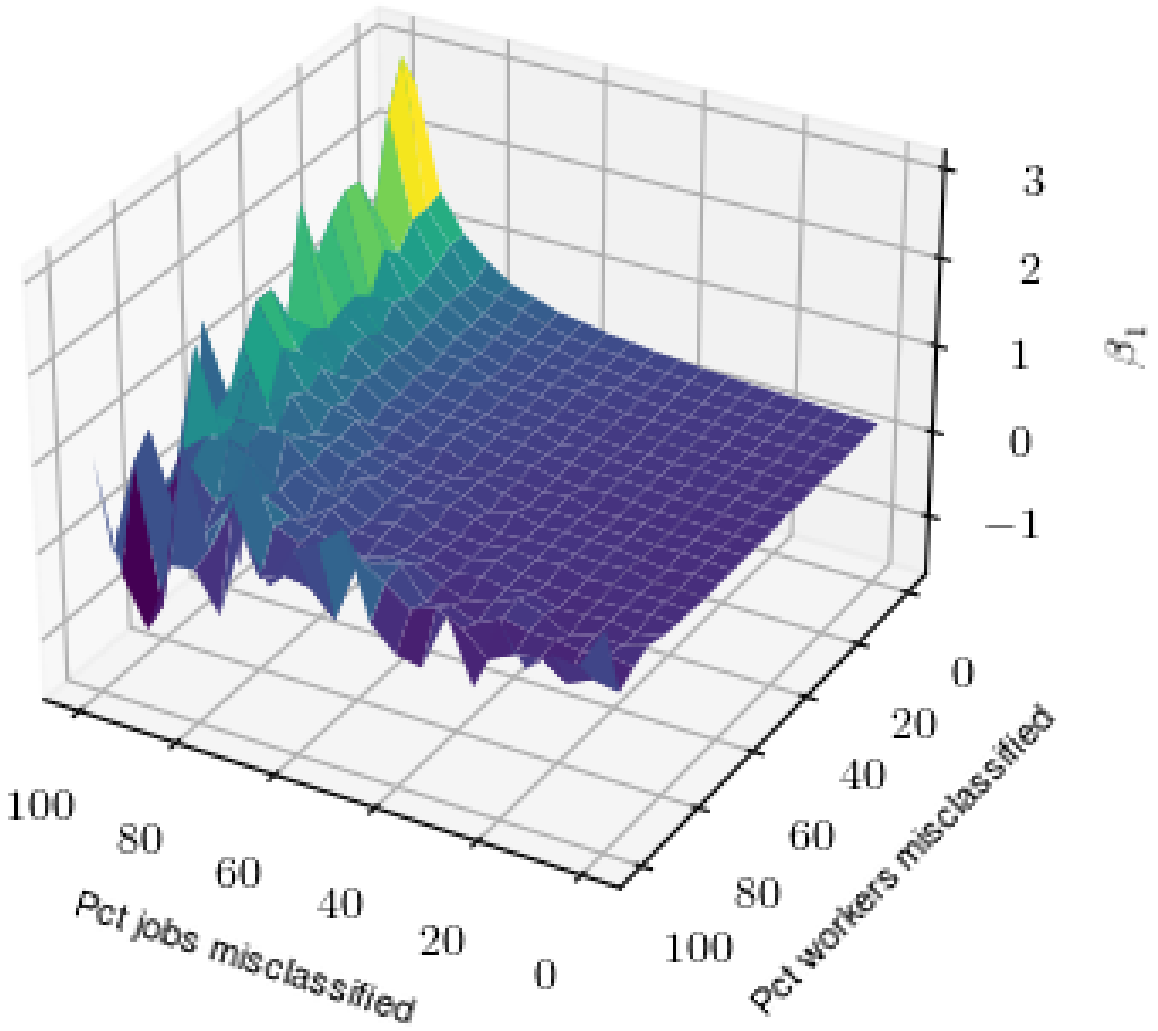
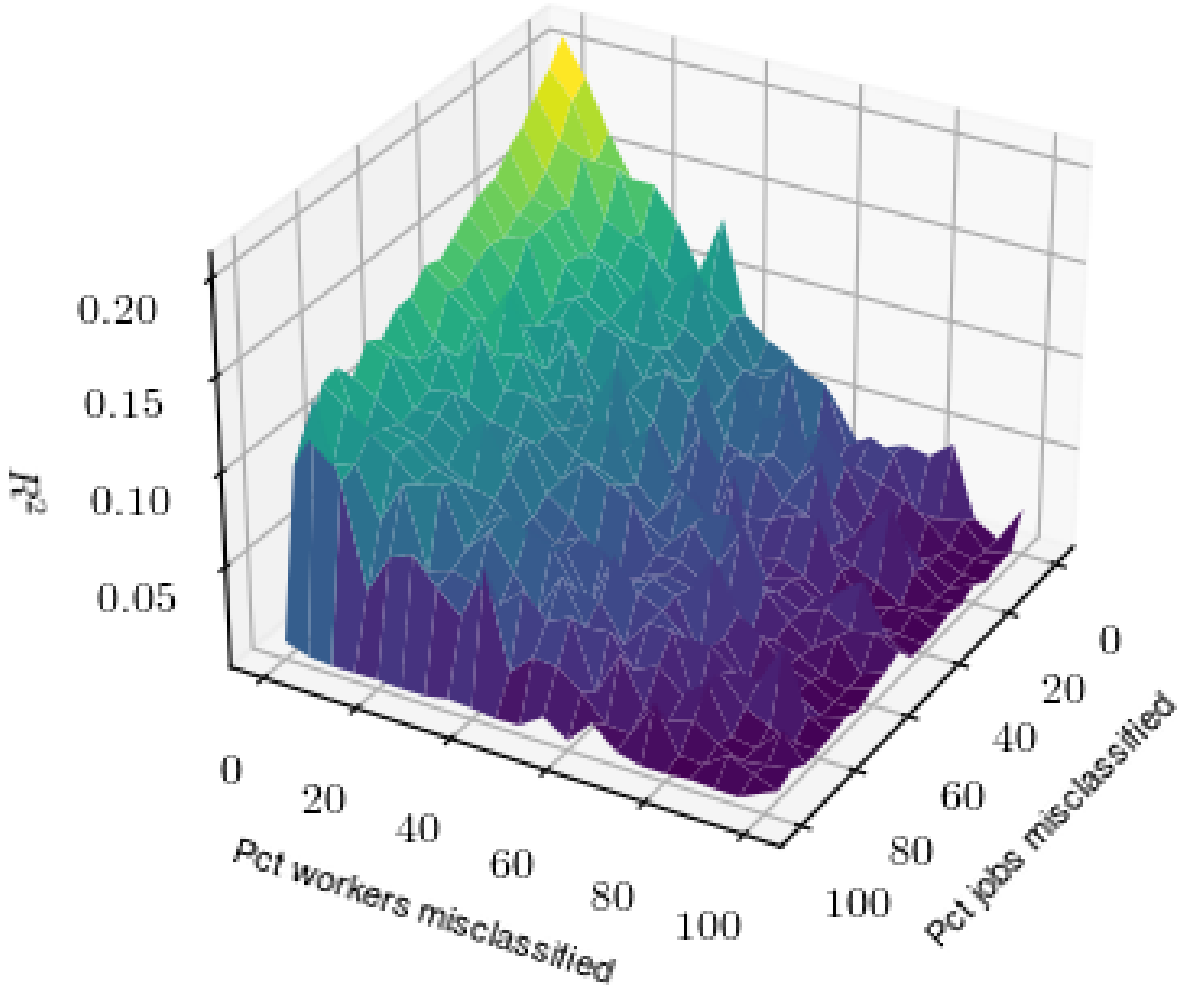


Figure 11: R^2 values with worker and job misclassification



H Proof that A_{ij} follows a Poisson distribution

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

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

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

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

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

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

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

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

Deriving the joint distribution:

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

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

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

If the summation term is equal to

$$\sum_{c_i=0}^{\infty} \frac{1}{(di - A_{ij})!} (1 - d_j P_{\iota\gamma})^{c_i - Aij} d_i^{c_i} = d_i^{A_{ij}} \exp(d_i(1 - d_j P_{\iota\gamma})) \tag{35}$$

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

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

Proving (35) is equivalent to proving the following equality:

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

Proof:

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

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

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

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

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

$= 1$

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

Therefore, we have

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

I Worker and firm fixed effects

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