



Cornell University
ILR School

Cornell University ILR School
DigitalCommons@ILR

Labor Dynamics Institute

Centers, Institutes, Programs

1-2017

Modeling Endogenous Mobility in Earnings Determination

John M. Abowd
Cornell University, jma7@cornell.edu

Kevin L. McKinney
U.S. Census Bureau

Ian M. Schmutte
University of Georgia

Follow this and additional works at: <https://digitalcommons.ilr.cornell.edu/ldi>

 Part of the [Labor Economics Commons](#), and the [Labor Relations Commons](#)

Thank you for downloading an article from DigitalCommons@ILR.

Support this valuable resource today!

This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Labor Dynamics Institute by an authorized administrator of DigitalCommons@ILR. For more information, please contact catherwood-dig@cornell.edu.

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact web-accessibility@cornell.edu for assistance.

Modeling Endogenous Mobility in Earnings Determination

Abstract

We evaluate the bias from endogenous job mobility in fixed-effects estimates of worker- and firm-specific earnings heterogeneity using longitudinally linked employer-employee data from the LEHD infrastructure file system of the U.S. Census Bureau. First, we propose two new residual diagnostic tests of the assumption that mobility is exogenous to unmodeled determinants of earnings. Both tests reject exogenous mobility. We relax the exogenous mobility assumptions by modeling the evolution of the matched data as an evolving bipartite graph using a Bayesian latent class framework. Our results suggest that endogenous mobility biases estimated firm effects toward zero. To assess validity, we match our estimates of the wage components to out-of-sample estimates of revenue per worker. The corrected estimates attribute much more of the variation in revenue per worker to variation in match quality and worker quality than the uncorrected estimates.

Keywords

earnings heterogeneity, mobility bias, latent class model, Markov Chain Monte Carlo

Disciplines

Labor Economics | Labor Relations

Comments

An earlier version of this paper can be found here: <http://digitalcommons.ilr.cornell.edu/ldi/23/>. This paper was previously released on 2016-05-19; the current version is more recent. The published version is available from Journal of Business and Economic Statistics at <http://dx.doi.org/10.1080/07350015.2017.1356727>

Replication code can be found at DOI:[10.5281/zenodo.376600](https://doi.org/10.5281/zenodo.376600) and our Github repository [endogenous-mobility-replication](https://github.com/abowd/endogenous-mobility-replication).

Abowd acknowledges direct support from NSF Grants SES-0339191, CNS-0627680, SES-0922005, TC-1012593, and SES-1131848. This research uses data from the Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

Modeling Endogenous Mobility in Earnings Determination

John M. Abowd

U.S. Census Bureau and Department of Economics, Cornell University
and

Kevin L. McKinney

Center for Economic Studies, U.S. Census Bureau
and

Ian M. Schmutte

Department of Economics, University of Georgia

January 2017

Abstract

We evaluate the bias from endogenous job mobility in fixed-effects estimates of worker- and firm-specific earnings heterogeneity using longitudinally linked employer-employee data from the LEHD infrastructure file system of the U.S. Census Bureau. First, we propose two new residual diagnostic tests of the assumption that mobility is exogenous to unmodeled determinants of earnings. Both tests reject exogenous mobility. We relax exogenous mobility by modeling the matched data as an evolving bipartite graph using a Bayesian latent-type framework. Our results suggest that allowing endogenous mobility increases the variation in earnings explained by individual heterogeneity and reduces the proportion due to employer and match effects. To assess external validity, we match our estimates of the wage components to out-of-sample estimates of revenue per worker. The mobility-bias corrected estimates attribute much more of the variation in revenue per worker to variation in match quality and worker quality than the uncorrected estimates.

Keywords: Earnings heterogeneity; Mobility bias; Latent-type model; Markov chain Monte Carlo

1 Introduction

We study the consequences of endogenous mobility for estimates of worker and firm effects on labor market earnings. Our starting point is the fixed-effects estimator for the statistical earnings model developed by Abowd et al. (1999, AKM, henceforth), which decomposes log earnings into components associated with unobserved worker and employer heterogeneity. Their statistical model assumes job-to-job mobility is exogenous with respect to the residual in their earnings equation. Structural interpretations of the estimated worker and firm effects are potentially useful as measures of unobserved skill and firm-specific earnings premia as long as mobility is exogenous. However, any bias from endogenous mobility will depend on how unmodeled factors that affect the evolution of wages are associated with the choices of individuals and firms to begin and end employment relationships. Our goal is to determine whether relaxing the exogenous mobility assumption has major implications for estimates of worker and employer-specific earnings heterogeneity.

We begin by clearly defining the exogenous mobility assumption entailed by the AKM log earnings model. The assumption of exogenous mobility implies that job-to-job mobility and job assignment depend only on time-invariant unobservable characteristics of workers and firms, along with available time-varying observable characteristics. While this allows for many forms of sorting, it precludes mobility driven by learning about new outside job opportunities (Woodcock 2008), mobility associated with learning about comparative advantage (Gibbons et al. 2005), or mobility based on idiosyncratic labor demand shocks (Helwege 1992). The exogenous mobility assumption is, therefore, subject to considerable skepticism among economists.

Despite this skepticism, there are no tests of exogenous mobility that can be computed from estimates made under the null hypothesis. We fill this gap by developing two new residual diagnostic tests of the exogenous mobility assumption, which we apply to longitudinally-linked employer-employee data from the U.S. Census Bureau's Longitudinal Employer Household Dynamics (LEHD) program. In both tests, the LEHD data reject the null hypothesis of exogenous mobility.

Rejecting the exogenous mobility assumption leaves open the question of whether and how endogenous mobility biases the estimates of worker- and firm-specific contributions to pay. To

address these questions, we develop a latent-type model of the data generating process that relaxes two key assumptions of the exogenous mobility model. First, our model allows for match effects that are correlated with worker and firm heterogeneity. Second, we allow both earnings and job mobility to be determined by latent types of workers, firms, and matches. We estimate the model using Bayesian methods that, because of the non-nested structure of the latent types, become computationally intensive. We exploit the network structure of the data and associated model restrictions to speed computation by using a graph coloring algorithm.

Our analysis of the uncorrected estimates from the AKM model derived under the assumption of exogenous mobility supports our modeling decisions. Our results suggest that: (1) all three sources of heterogeneity contribute to earnings variation although worker and employer effects are more important than match effects in the structural estimates as compared to the AKM estimates; (2) structural match effects are negatively correlated with structural worker and firm effects (impossible under exogenous mobility); (3) there is almost no evidence of positive assortative matching in the structural worker and firm effects; and (4) in the steady-state, the endogenous mobility model exhibits the same wage compression due to negative correlation of the worker and firm effects with the match effect as seen in the simple correlations.

We go on to validate our estimates using data on firm revenues. Using the uncorrected AKM estimates, we find a strong association of firm revenue with firm effects, but only a very weak association of revenue with firm-level averages of worker effects and within-match mean residuals. Using the endogenous-mobility corrected estimates, we find revenue is strongly associated with all three components of earnings heterogeneity. The latter pattern is consistent with models in which worker and match effects in earnings represent compensation for productive attributes.

Our analysis and methods are relevant in applications that use estimates of worker and firm effects as measures of human capital and firm-specific earnings premia. Examples include the analysis of firm productivity (Iranzo et al. 2008), inter-industry earnings differentials (Abowd et al. 2012), exporter earnings differentials (Krishna et al. 2014), job referral networks (Schmutte 2015), and earnings inequality (Card et al. 2013). Endogenous assignment also affects the estimation of neighborhood effects on earnings (Combes et al. 2008), and value-added models of student test scores (Rothstein 2010; Koedel and Betts 2010; Kramarz et al. 2014).

2 Background and Motivation

2.1 The AKM Model for Labor Market Earnings

AKM proposed a framework for modeling the logarithm of individual labor earnings in matched employer-employee data that allows for arbitrary heterogeneity across workers and across employers. In the AKM model, the log earnings of worker i in period t is

$$\ln w_{it} = X_{it}\beta + \theta_i + \psi_{J(i,t)} + \varepsilon_{it}. \quad (1)$$

The log earnings depend on observable characteristics through $X_{it}\beta$, unobservable individual-level characteristics, θ_i , and an unobservable employer-specific component, $\psi_{J(i,t)}$, where $J(i,t)$ is a function that maps worker-year observations to their unique employer in that year. The method for selecting the unique employer is elaborated in Section 5.1.1. This model applies to data that include observations on I individual workers and J employers. The data are observed over T periods and in any period t , there are a total of N_t observations. The full sample includes N worker-period observations.

In matrix notation, the earnings model (1) is

$$\ln w = X\beta + D\theta + F\psi + \varepsilon \quad (2)$$

where $\ln w$ is the $[N \times 1]$ stacked vector of log earnings outcomes $\ln w_{it}$, X is the $[N \times k]$ design matrix of observable time-varying characteristics. In the analysis that follows, and in the empirical work, $\ln w$ and X are measured as deviations from their overall means and we suppress the constant term. The matrix D is the $[N \times I]$ design for the individual effects; F is the $[N \times J]$ design matrix for the employer effects (non-employment is suppressed here). The unknown fixed effects to be estimated, $\left[\beta^T \quad \theta^T \quad \psi^T \right]^T$, have dimension $[k \times 1]$, $[I \times 1]$, and $[J \times 1]$ associated with each of the design matrices.

Identification and the Exogenous Mobility Assumption

Identification of the parameter vector $\begin{bmatrix} \beta^T & \theta^T & \psi^T \end{bmatrix}^T$ in the statistical model requires the following orthogonality conditions:

$$E[X^T \varepsilon] = 0; E[D^T \varepsilon] = 0; E[F^T \varepsilon] = 0. \quad (3)$$

As long as the matrix of data moments has full rank, these conditional moment restrictions yield an unbiased estimator for the full parameter vector, including the individual and employer effects.

The assumptions in (3) that $E[D^T \varepsilon] = 0$ and $E[F^T \varepsilon] = 0$ are particularly problematic. They imply that there is no correlation between the earnings residual and an individual's decision to enter or exit the labor market, and that there is no correlation between the residual and the assignment of workers to employers. These assumptions do not have a clear behavioral foundation, but they follow from the stronger assumption that mobility and assignment are independent of the earnings residual.

Specifically, we define exogenous mobility by the assumptions:

$$\begin{aligned} E[\varepsilon|X] &= 0 \\ \Pr[D, F|X, \varepsilon] &= \Pr[D, F|X]. \end{aligned} \quad (4)$$

If the exogenous mobility assumptions are satisfied, it is clear that $E[\varepsilon|X, D, F] = 0$. The orthogonality conditions necessary for identification follow. See Abowd et al. (2002) for the method of ensuring that D and F have full column rank.

Exogenous mobility requires that a worker's employment history is completely independent of the idiosyncratic part of earnings captured in ε . Specifically, knowledge of the entire history of earnings residuals does not convey any information that would help predict job assignment or worker entry and exit. The exogenous mobility assumption is, thus, equivalent to assuming that all assignments are pre-determined at birth given full knowledge of X, D, F and $\begin{bmatrix} \beta^T & \theta^T & \psi^T \end{bmatrix}^T$ (Rothstein 2010).

3 Empirical Model

To relax the assumption of exogenous mobility, we develop a model in which a match-specific component of the earnings residual predicts the movement of workers between employers. The huge number of workers and employers render the problem of predicting job mobility extremely challenging. To make progress, we use a latent-type framework. The three populations of interest – workers, employers, and job matches – are associated with latent heterogeneity types that affect earnings, mobility and job assignment. Our model allows for arbitrary correlation between worker and employer types on observed matches, and also allows for sorting by comparative advantage by allowing arbitrary correlation between job match quality and worker and employer attributes.

3.1 Model Setup

3.1.1 Population Heterogeneity

The agents are workers, indexed by $i \in \{1 \dots I\}$ and employers, indexed by $j \in \{0 \dots J\}$. By convention, when a worker is not employed, we say he is assigned to employer $j = 0$. On entry to the labor market, a worker i samples his type from one of L latent ability types, $a_i \in \mathcal{A}$. Likewise, each employer, except $j = 0$, samples her type from one of M latent types, denoted $b_j \in \mathcal{B}$. Again, by convention, the non-employment state is associated with its own latent type, b_0 ; hence, there are $M + 1$ employer types counting non-employment.

In our empirical application, we have access to data on the complete population of workers and employers. Each potential worker-employer match (a job) has an associated latent heterogeneity component that affects both earnings and mobility: $k_{ij} \in \mathcal{K}$, where \mathcal{K} has cardinality Q . To make the subsequent formulas easier to interpret, we represent the elements of \mathcal{A} , \mathcal{B} and \mathcal{K} as rows from the identity matrices I_L , I_{M+1} and I_Q , respectively. For instance, if the number of latent worker types, $L = 2$, then $\mathcal{A} = \{(1, 0), (0, 1)\}$.

We assume workers and employers sample their latent ability and productivity types independently from multinomial distributions with parameters π_a, π_b . However, the distribution of match quality is not independent of worker or employer type. The probability that the latent type of the match between worker i and employer j is k is $\Pr(k_{ij} = k | a_i = a, b_j = b) = \pi_{k|ab}$. This spec-

ification allows for independent match effects as a special case. If match quality is independent of worker and employer heterogeneity, then AKM estimates of the worker and firm effects are unbiased even in the presence of endogenous mobility.

3.1.2 Earnings Determination

The log of earnings on any match is given by the following generalization of the AKM model

$$\ln w_{ijt} = \alpha + X_{ijt}\beta + a_i\theta + b_j\psi + k_{ij}\mu + \varepsilon_{it}. \quad (5)$$

The vector X_{ijt} includes observable time-varying characteristics. In practice, we use the same vector of observable characteristics that are used for our estimates of the AKM model. We describe the vector of control variables in Section 5.2. The vectors θ , ψ and μ are parameters with dimension $L \times 1$, $M \times 1$, and $Q \times 1$, respectively, that describe the effect on the level of log earnings associated with membership in the various heterogeneity types. We take ε to be normal with mean 0 and variance σ^2 , independent and identically distributed across individuals and over time.

3.1.3 Mobility Model

The definition of exogenous mobility in Equation 4 implies that the assignment of workers to employers and job durations should not depend on any function of the earnings residual. Our diagnostic analysis in Section 5.2 suggests job mobility and assignment are predicted by a match-specific component of earnings that may be correlated with worker and firm heterogeneity. We therefore develop a model of endogenous mobility with these features.

The separation indicator $s_{it} = 1$, if i separates from his current job at the end of period t , and $s_{it} = 0$, otherwise. Recall $J(i, t)$ is the index function that returns the identifier of the firm in which i is employed in period t . The probability of separation depends flexibly on match quality:

$$\Pr [s_{it} = 1 | k_{iJ(i,t)}] = f_{se}(a_i, b_{J(i,t)}, k_{iJ(i,t)}; \gamma) \equiv \gamma_{abk} \quad (6)$$

where $0 \leq \gamma_{abk} \leq 1$.

Furthermore, conditional on separation, the latent type of the next employer depends on the

quality of the current match, in addition to the productivity of the current employer and the ability of the worker:

$$\Pr [b_{J(i,t+1)} | a_i, b_{J(i,t)}, k_{i,J(i,t)}] = f_{tr} (a_i, b_{J(i,t)}, k_{i,J(i,t)}; \delta) \equiv \delta_{abk} \in \Delta^{M+1} \quad (7)$$

where $\delta_{abk} \equiv [\delta_{0|abk}, \dots, \delta_{M|abk}]$ is a $1 \times (M + 1)$ vector of transition probabilities, Δ^{M+1} is the unit simplex, and $J(i, 0) = 0$ for all i . The transition probabilities are indexed by all of the latent heterogeneity in the model. Within a heterogeneity type, the identity of the precise employer is selected completely at random.

3.2 Interpretation

It is instructive to discuss how exogenous mobility can be expressed in terms of restrictions on the model of Section 3.1. Specifically, exogenous mobility requires that separation probabilities and job assignments be independent of match quality. These assumptions are exactly the restrictions $\gamma_{abk} = \gamma_{ab}$ and $\delta_{abk} = \delta_{ab}$. For our purposes, it is also important to understand how these assumptions interact with assumptions on the match effect, μ , in the earnings equation. The conventional AKM model assumes no match effect (or, equivalently that the match effect is independent of worker and firm effects). If we make this assumption as well, so that $\pi_{k|ab} = \pi_k$ then our model reduces to a latent type version of the AKM model.

Identification in the endogenous mobility model is based on using the relational structure of the data to predict job matches. The labor market is an evolving graph of connections between workers and their employers. At time t , let the set of identifiers for all I individuals who work in one of the $J + 1$ employers (including non-employment), $A(t)$, and the set of $J + 1$ employers, $E(t)$, be arranged in a bipartite graph where $A(t)$ and $E(t)$ are the two (disjoint) vertex (or node) sets. There is a link between $i \in A(t)$ and $j \in E(t)$ if and only if i is employed by j at date t . The totality of the links (jobs) active at date t , excluding non-employment, can be represented as an $I \times J$ matrix $B(t)$, which is the upper right-hand block of the full adjacency matrix for the bipartite graph, with i on the rows and j on the columns, and is sometimes called the biadjacency matrix.

The collection of labor market relationships at time t , summarized by the adjacency matrix $B(t)$, we call the *realized employment network*. The observed labor market data are snapshots of the market at points in time, $B(t_1), \dots, B(t_T)$, where T is the total number of available time periods. Our adjacency matrix representation can be directly related to the AKM framework. When the data are sorted first by time, t , and then by workers, i , the design matrix of employer effects is

$$F = \begin{bmatrix} B(1) \\ B(2) \\ \vdots \\ B(T) \end{bmatrix}$$

where $B(t)$ is the adjacency matrix from the bipartite labor market graph.

Each adjacency matrix, $B(t)$, describes which outcomes (job matches) were observed at each point in time from the collection of $I \times J$ potential outcomes at each moment of time, again excluding non-employment. The potential outcomes are given by the structural earnings equation (5). Under exogenous mobility, potential outcomes are selected for observation conditionally-at-random, given X . In this sense, the problem of endogenous mobility is a sample selection problem.

To address these selection biases, our model groups together workers and firms with similar mobility, earnings, and turnover patterns by assigning them to latent types. The model exploits that, in expectation, workers of the same type matched to firms of the same type have, ex ante, the same probability of separation and, conditional on separation, the same expected destination.

Our application and proposed procedure are, therefore, also related to stochastic block models, modularity maximization, and other methods for the detection of “communities” of nodes in complex social and economic networks. Our main innovation is the use of both node and edge characteristics in predicting the matches (Hoff et al. 2002; Newman and Leicht 2007; Schmutte 2014).

In this class of models, for a fixed finite dimensional heterogeneity space, the posterior distributions of the parameters characterizing unobserved heterogeneity and of the earnings and mobility equations are all proper. Therefore, the Gibbs sampler we will use is well-behaved, producing samples from the joint distribution of the latent data and model parameters. Elements of the likeli-

hood function, such as the homoscedastic normal kernel, could be replaced with alternative, finite parameterizations of the earnings process. Elements of the first-order discrete Markov job transition process could be generalized to higher finite orders. In Section E.4, we consider the consequences of changing the dimension of the latent worker, firm and match heterogeneity, but we restrict attention to relatively low-dimensional representations. Thus, our modeling decisions stay entirely within the realm of feasible finite-dimensional models for which our posterior sampler is well-behaved. We are unaware of any nonparametric Bayesian analyses of models with the same network complexity as ours—latent factors in both node types and on the edges.

Bonhomme et al. (2015) also develop a model in which workers and firms have latent unobserved types that affect earnings and mobility. Their focus is on accounting for the possible presence of nonseparable worker and firm heterogeneity in the earnings equation. By contrast, the goal of our model is to consider the biases that can arise from endogenous mobility in the presence of an omitted correlated match effect. Our model in principle nests the case where worker and firm heterogeneity are fully non-separable. Furthermore, it allows for the possibility, supported empirically by the evidence in Figure 3, that non-separability, or the role of match heterogeneity, varies over the distribution of worker and firm types.

3.3 Likelihood Function

The observed data, y_{it} , consist of labor earnings, observable time-varying characteristics, separations, and employer identifiers:

$$y_{it} = [\ln w_{iJ(i,t)t}, X_{it}, s_{it}, i, J(i, t), J(i, t + 1)] \text{ for } i = 1, \dots, I \text{ and } t = 1, \dots, T. \quad (8)$$

The latent data vector, Z , consists of the heterogeneity types:

$$Z = [a_1, \dots, a_I, b_1, \dots, b_J, k_{11}, k_{12}, \dots, k_{1J}, k_{21}, \dots, k_{IJ}]. \quad (9)$$

Finally, the complete parameter vector is

$$\rho^T = [\alpha, \beta^T, \theta^T, \psi^T, \mu^T, \sigma, \gamma, \delta, \pi_a, \pi_b, \pi_{k|ab}], \rho \in \Theta. \quad (10)$$

The observed data matrix is denoted Y . The likelihood function for the parameters is the joint distribution of observed and latent data (Y, Z) :

$$\mathcal{L}(\rho|Y, Z) \propto \prod_{i=1}^I \left\{ \begin{aligned} & \times \prod_{t=1}^{T-1} \left[1 - \gamma_{\langle a_i \rangle \langle b_{J(i,t)} \rangle \langle k_{iJ(i,t)} \rangle} \right]^{1-s_{it}} \left[\gamma_{\langle a_i \rangle \langle b_{J(i,t)} \rangle \langle k_{iJ(i,t)} \rangle} \right]^{s_{it}} \\ & \times \prod_{t=1}^{T-1} \left[\delta_{\langle b_{J(i,t+1)} \rangle | \langle a_i \rangle \langle b_{J(i,t)} \rangle \langle k_{iJ(i,t)} \rangle} \right]^{s_{it}} \end{aligned} \right\} \\ \times \prod_{i=1}^I \prod_{j=1}^J \left[\left(\prod_{\ell=1}^L (\pi_{a\ell})^{a_{i\ell}} (\pi_{bm})^{b_{jm}} (\pi_{q|\ell m})^{k_{ijq}} \right) \right] \quad (11)$$

where the notation $\pi_{a\ell}$ denotes the ℓ^{th} element of π_a (similarly for π_{bm} , etc.) and $\langle x \rangle$ means the index of the non-zero element of the indicator vector x .

The likelihood function factors into a part due to the observed data conditional on the latent data, and the latent data conditional on the parameters. The observed-data likelihood conditional on the latent data factors further into separate contributions from the earnings and the mobility processes. The mobility process is Markov, and conditionally independent of the earnings realizations once we know the latent types of the workers, firms and matches. We assume that the matches initially observed are exogenous.

The power of the model comes from the predictive equation for the latent data Z – the unobserved types associated with worker, employer, and match heterogeneity. Given the observed data and the parameters, the posterior predictive distribution of Z is computed as the complete-data likelihood divided by the observed-data likelihood. The observed-data likelihood is calculated by integrating out the latent data. We describe our estimation procedure in detail in Section 4.

4 Estimation Method

Our empirical approach is Bayesian, but standard estimation techniques are not effective because the worker and employer effects are not nested. We estimate the model by adapting the Gibbs sampler for finite mixture models as developed in Tanner (1996) and Diebolt and Robert (1994) to our model, allowing for multiple overlapping levels of correlation across observations. As in

the data augmentation algorithm, we iterate between sampling from the posterior distribution of the parameters given the complete data – both observed and latent data – and sampling from the posterior predictive distribution of the latent data – the unobserved worker, employer, and match types – given the model parameters.

Below, we derive the posterior distribution of the model parameters given the latent data. The derivation is standard, as is the method for sampling. The more challenging task, both analytically and computationally, is deriving, and sampling from, the posterior predictive distribution for unobserved types. Under the model, the posterior distribution for the latent type depends on the relative likelihood contribution of a particular entity (worker, employer or match) under each latent type. Generically, the likelihood contribution of a worker depends not only on his own ability type, but also on the productivity type of the employer he works for, and the match quality type.

We exploit the relational structure of the data together with conditional independence assumptions implied by the model to facilitate computation. We update the worker effects first, then the employer effects, then the match effects. Furthermore, the conditional independence assumptions in the model imply that the latent types of the workers can be updated simultaneously (in parallel). The same is true of the match heterogeneity types.

For employer heterogeneity, the situation is more complex. Because the probability of assignment to a new employer is a function of the previous employer, the likelihood contribution of a given employer is not independent of the employers to which it is directly connected through the realized mobility network. Therefore, the latent type of any employer is not independent of the type of its network neighbors. Without further analysis, this requires that the employer types be updated sequentially, which is very time consuming.

We use the network structure of the data together with our modeling assumptions to parallelize the computation. Specifically, under the model for the observed and latent data, the probability that a firm is of a particular type is independent of the types of all firms with which it is not connected through a direct job-to-job transition; that is, firms that do not have a degree-one network connection. We apply a graph coloring algorithm to the employer projection of the realized mobility network to partition employer nodes into disjoint groups, based on this conditional independence assumption, that can be updated in parallel. The effectiveness of this technique therefore depends

on both the conditional independence assumption in the model and the network structure of the data. If all firms were degree-one connected to one another through job-to-job flows, this method would not provide any parallelization. Instead, we find that firms can be separated into just 24 separate groups within which no two firms are connected by a direct job-to-job transition.

4.1 Prior and Posterior distributions

The prior on the parameter vector $\rho^T = [\alpha, \beta^T, \theta^T, \psi^T, \mu^T, \sigma, \gamma, \delta, \pi_a, \pi_b, \pi_{k|ab}]$ is the product of priors on its component terms. Each vector of probabilities, $\gamma, \delta, \pi_a, \pi_b, \pi_{k|ab}$, has a Dirichlet distribution with prior sample size equal to one. Each element of the Dirichlet parameter vector is given by the inverse of the dimension of the probability vector.

We choose conjugate priors for the earnings and mobility model parameters. Conditional on the population type probabilities, $\pi_a, \pi_b, \pi_{k|ab}$, the coefficients in the log earnings equation, $(\alpha, \beta^T, \theta^T, \psi^T, \mu^T)$ have uninformative normal prior distributions. The variance parameter, σ , has the inverted gamma prior $\text{IG}(\nu_0, s_0)$. We set $\nu_0 = 1$ and $s_0 = 1$.

We also constrain the population probability-weighted average earnings heterogeneity effects to be zero. That is, $\pi_a^T \theta = \pi_b^T \psi = \pi_{k|ab(\ell m)}^T \mu = 0$ for all ℓ, m where $\pi_{k|ab(\ell m)} \equiv \Pr[k_{ij} = k | a_i = \ell, b_j = m]$. This assumption highlights the inherent sample selection problem. If all workers were observed in all matches in all periods, there would be no endogenous mobility bias because all counterfactuals would be observed, not latent. With the preceding assumptions, we derive the posterior distributions for each parameter in Appendix C.

4.2 Gibbs Sampler

We start the Gibbs sampler with initial values for the parameter vector and latent data, $\rho^{(0)}, Z^{(0)}$. To update the parameter vector, we sample from the posterior distributions defined in Appendix C. We must still define the posterior distribution for the latent data, Z , given the observed data and parameters. To update the ability types for the workers, we sample from a multinomial with probability that worker i falls in the ℓ^{th} type:

$$\Pr(a_i = \ell | a_{-i}, b, k, Y, \rho) = \frac{\Pr(a_{-i}, b, k, Y | \rho, a_i = \ell) \Pr(a_i = \ell)}{\Pr(a_{-i}, b, k, Y | \rho)}$$

$$= \frac{\pi_{a\ell} \Pr(a_{-i}, b, k, Y | \rho, a_i = \ell)}{\sum_{\ell'=1}^L [\pi_{a\ell'} \Pr(a_{-i}, b, k, Y | \rho, a_i = \ell')]}. \quad (12)$$

where a_{-i} represents the types of all workers other than i . To calculate equation (12) requires computing the likelihood contribution of i when assigned to each of the L ability types. In our model, the posterior probability of a_i is independent of a_{-i} , conditional on the rest of the data (latent and observed): $\Pr(a_i = \ell | a_{-i}, b, k, Y, \rho) = \Pr(a_i = \ell | b, k, Y, \rho)$. This conditional independence allows us to speed computation by updating the latent types of each worker in parallel. The proof is a straightforward consequence of the conditional independence across workers in the likelihood function for the complete data.

The posterior predictive distribution for the latent match quality is $\Pr(k_s = q | a, b, k_{-s}, Y, \rho) = \Pr(k_s = q | a, b, Y, \rho)$, which likewise follows from the conditional independence assumptions in the model. Hence, for a given type of workers and employers, (a, b) , the latent quality of each match is conditionally independent from the others. We exploit the conditional independence by parallelizing these updates as well.

The posterior distribution for employer types exhibits a conditional dependence that is not present for workers or matches. When a worker changes jobs, the latent type of the employer for the successor job depends on the the latent type of the employer on the origin job. Therefore, the posterior probability that a firm is of a particular type depends directly on the types of firms to which it is connected through the realized mobility network.

The posterior distribution is

$$\Pr(b_j = m | a, b_{-j}, k, Y, \rho) = \Pr(b_j = m | a, b_{N(j)}, k, Y, \rho) \quad (13)$$

where $b_{N(j)}$ denotes the latent types of the employers in $N(j)$, the set of neighbors of j (in the employer projection of the realized mobility network). We use this result to define a partition of employers into groups that can be updated in parallel. The details of our graph-coloring algorithm appear in Appendix E.1.

5 Data

We implement the model empirically using matched employer-employee data from the LEHD program of the U.S. Census Bureau. Our analysis compares estimates from the standard AKM decomposition with estimates from the model described in Section 3. This comparison indicates that relaxing the exogenous mobility assumption may have a large effect on estimated worker, employer, and match effects. To validate our model, we bring in data on firm revenue, which is not part of the LEHD infrastructure file system. This section briefly describes the sources of data and how we prepared the research files.

5.1 Data Sources

5.1.1 LEHD Analysis Population

Our analysis population is the universe of dominant job records from the LEHD data for the states of Indiana, Illinois, and Wisconsin for the years 1999–2003. If a worker has earnings reported from multiple employers during the year, the dominant job is the one with highest earnings. We restrict attention to workers who are never younger than 18 nor older than 70. There are 60,123,894 person-year observations in the population covering 15,998,626 workers and 712,494 firms. Details of the data sources and variable preparation are described in Appendix E.3.

We estimate the complete AKM decomposition on this analysis population, and use the estimates to perform residual diagnostic tests of the exogenous mobility assumption. We describe these tests in Section 5.2. We also use our full population estimates of the AKM parameters, θ_{AKM} and ψ_{AKM} to create initial types of workers, firms, and matches for the Gibbs sampler. For each match, we construct an AKM orthogonal match effect, μ_{AKM} , as the average residual during the match. Next, we construct the deciles of θ_{AKM} , ψ_{AKM} , and μ_{AKM} within their respective populations (across workers, across firms, and across matches).

5.1.2 LEHD Sample for Structural Estimation

Estimation of the structural model on the full analysis population is computationally infeasible. We therefore draw a 0.5% simple random sample of workers from the analysis population, re-

taining their full employment histories. The final analysis sample includes 395,930 Person-year observations (including years spent in non-employment) that cover 79,186 Persons, 60,589 Firms, and 133,870 Matches. We assign each observation in the analysis sample the appropriate θ -decile, ψ -decile, and μ -decile based on the AKM estimates to use as starting values for the latent types of workers, employers and matches.

5.1.3 Firm Revenue Data

We match our analysis sample with firm-level data on total revenue per worker. The firm-level revenue data is built up from the 2002 Economic Census. For the economic census, revenue is defined as “the total sales, shipments, receipts, revenue or business done by domestic establishments (excludes foreign subsidiaries) within the scope of the economic census.” (U.S. Census Bureau 2006, p. A-2). Sector-specific variations are elaborated in U.S. Census Bureau (2006, pp. A-2-A-6). The population for the revenue data is all establishments that appear in the Census Employer Business Register. The economic census collects sales data from a sample of establishments. The sampling is based on industry and establishment size. For non-sampled establishments, which are missing conditionally at random (i.e., ignorably missing), the missing revenue data are multiply imputed using a model that conditions on all frame variables. The result is a dataset that contains the universe of all business establishments with complete data on revenue and employment. Next, the establishments in the LEHD data are matched to establishments in the business register. The establishment identifiers in the LEHD are distinct from the establishment identifiers in the business register, necessitating the use of a statistical matching procedure. For firms with multiple establishments we sum revenue and employment across all establishments and compute revenue per worker directly.

5.2 Estimation and Diagnostic Analysis of the AKM Model

We estimate the AKM model from equation (2) using the method described in Abowd et al. (2002). When estimating the AKM model and the structural endogenous mobility model, the vector of time-varying controls, X_{it} contains a quartic in age interacted with gender, race, and ethnicity, year controls, and a detailed set of controls for attachment to the job over the year. The latter

controls, for job attachment, are built from the quarterly earnings records. We include them to address the problem that predicted earnings will otherwise be too large mechanically in years of job transition. A complete discussion of the control variables and their role in estimation appears in Appendix E.3.

In our estimate of the AKM decomposition on the LEHD analysis population, firm effects explain 15 percent of the variation in log earnings. Fixed worker effects explain 32 percent. Observable, time-varying characteristics explain an additional 46 percent of earnings variation. When we restrict our analysis to the 0.5% sample used in structural estimation, the shares of earnings explained by each component are identical. Table OA1 reports the complete matrix of estimated covariances for our analysis sample, along with the analogous covariances for the endogenous mobility model.

5.2.1 Testing the Exogenous Mobility Assumption

We develop two tests that exploit the exogenous mobility assumption that current earnings residuals should not be predictive of future employer assignments. Both tests strongly reject the null hypothesis of exogenous mobility for these data. We describe the procedure for calculating both test statistics in detail in Appendix B.

The first test, the *match effects test*, checks whether the firm effect of a worker's future employers are independent of the average residual in the current job. The match effects test yields a test statistic, $X^2 = 1,169,205$, that is distributed chi-square with 8,991 degrees of freedom. Using conventional criteria, this test has a p-value less than 10^{-6} .

The second test, the *productive workforce test*, checks whether the average worker effect of future employees of a particular employer are predictive of the residuals on that firm's current period earnings payments. The productive workforce test has a test statistic $X^2 = 34,950$ that is distributed chi-square with 900 degrees of freedom. Again, using conventional criteria, this test has a p-value less than 10^{-6} .

As discussed in Andrews et al. (2008), the estimated firm effects contain a linear combination of all residuals for all workers in all time periods. In short samples, this could induce a spurious correlation between current residuals and future firm effects, even if the the assumptions of ex-

ogenous mobility hold. We recomputed the match effects test using a hold-out sample to compute the average match-specific residual and obtain a test statistic of 619,539. We conclude that the qualitative features of our analysis of residuals from the AKM decomposition are not driven by spurious correlation bias.

5.2.2 Evidence for Match Effects in the AKM Results

Figures 1a and 1b display the average AKM residual within cells defined by unique pairs of worker effect - firm effect deciles. Figure 1a displays this information as a three-dimensional bar graph, akin to Card et al. (2013). Unlike the German data analyzed by Card et al. (2013), the LEHD sample exhibits more variation in the average match effect across cells. Therefore, we report the same information in Figure 1b as a grouped bar graph. The data are grouped by worker effect deciles along the horizontal axis. Within each group we plot bars for each firm effect decile, given the worker effect decile, whose heights correspond to the average match effect within each. Figure 1b is thus a “flattened” version of the same information displayed in Figure 1a. While the average residuals are less than 2 log points in many cells, their magnitudes are sometimes larger and vary considerably across worker deciles and across firm deciles. In particular, the residual earnings of low earnings workers vary considerably with the type of firm. Likewise, residual earnings on jobs in low-earnings firms vary considerably with the type of worker employed. As a diagnostic tool, Figure 1a supports our inclusion of correlated match effects in the structural model.

This finding relates our work to the theoretical literature dealing with the link between the AKM earnings decomposition and assignment models of the labor market, particularly Eeckhout and Kircher (2011) and Shimer (2005). These papers demonstrate that person and firm effects are not sufficient to identify sorting on unobserved worker ability and firm productivity. These assignment models often, though not necessarily, imply that a match effect has been omitted from the structural earnings equation. Assignment models also imply that the estimated person and firm effects are complicated transformations of parameters describing the latent productivity and ability of the underlying populations and workers and employers together with the equilibrium earnings and assignment process. It is therefore very difficult to say whether the patterns observed in Figure 1a and the remainder of this paper are consistent with any specific assignment model. Abowd

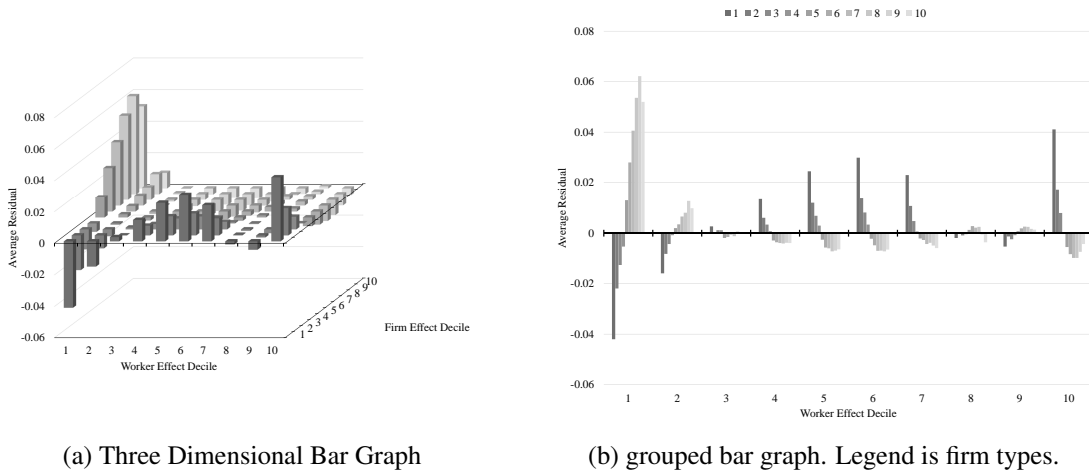


Figure 1: Mean residual from the AKM decomposition by Worker/Firm effect deciles.

et al. (2015) explore this matter formally with respect to the coordination friction model in Shimer (2005).

5.2.3 Evidence on the Exogenous Mobility Assumption

The exogenous mobility assumption in equation 4 implies, among other things, that for workers who change jobs, the match effect (average residual) on the origin job should not help predict the destination job. A straightforward corollary, following Bayes' Rule, is that the average match effect for workers changing jobs is independent of the type of firm to which they are transiting. Knowledge of the future firm type should be uninformative about the current match effect after controlling for the type of the worker and the type of origin firm under exogenous mobility.

We assess this corollary implication graphically in Figure 2. The figure displays a grouped bar graph that shows the average residual within cells defined by origin firm–destination firm pairs. The figure is organized with origin firm types along the horizontal axis. Within each origin firm type are displayed 10 bars whose height represents the average residual on the job from which the worker separates, for each of the ten destination firm types.

If the assumption of exogenous mobility were valid, the average match effect on the origin job should not help predict destination firm type. If this were the case, then Figure 3 would show a repeating pattern across all destination firm types. Instead, there is a visually evident correlation

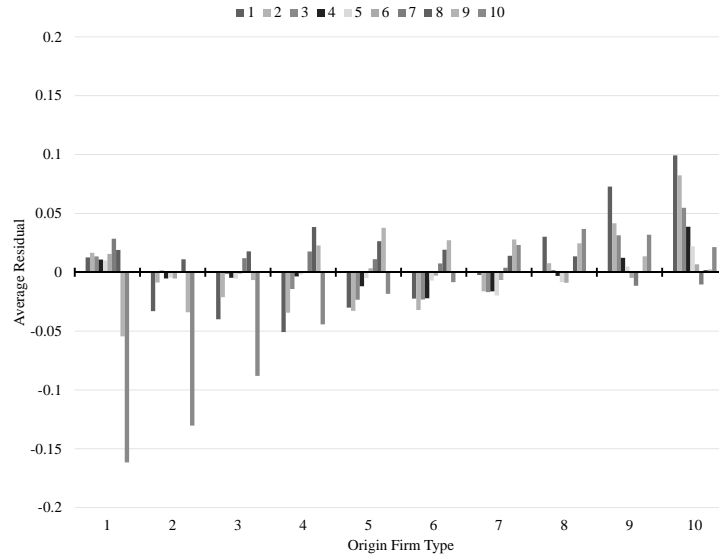


Figure 2: Mean AKM residual within origin/destination firm effect decile. Legend is for destination firm types.

between destination firm and the average match effect. For example, among workers whose origin firm has a high estimated firm effect, movements to low-earnings firms are predicted by high average residuals on the origin job. Likewise, for workers employed initially in low-earnings firms, movement to high-wage firms is predicted by a low average residual on the origin job.

Figure [OA1](#) offers a different view on these results, by plotting the average change in residual for workers as they change jobs. Again, if mobility is exogenous, the change in residual should be zero in expectation, and not predictive of the origin or destination firm type. Instead, we observe that job transitions from the lowest decile to the highest decile of firm effects are associated with large positive changes in residual earnings. Likewise, job transitions from the highest deciles to lower deciles are associated with negative changes in residual earnings.

Finally, the data show some evidence of earnings compression across worker types, and across firms within worker types. The average match effects are larger, and more consistently negative, as we move from Worker Type 1 (low-earnings workers) to Worker Type 10 (high-wage workers). Focusing on Figure [3c](#), within a specific destination firm type, there is some evidence that the average match quality is higher on origin jobs in low-earnings firms and lower on origin jobs in high-wage firms. A similar pattern is exhibited across firm types in Figure [1b](#). Altogether,

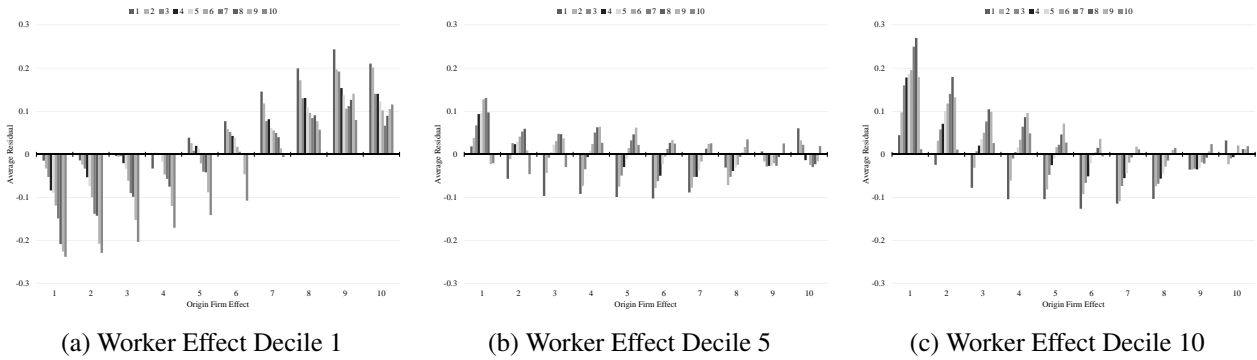


Figure 3: Mean AKM residual within origin/destination firm effect decile, disaggregated by worker effect decile. Legend is for destination firm types.

these plots reinforce the residual diagnostic tests in Section 5.2.1 and help illustrate what form endogenous mobility bias may take.

What sort of model might produce the observed patterns? For the high-type workers, the observed pattern is consistent with selection: workers only leave jobs with high match effects when moving to high-paying firms. For the low-type workers, the observed pattern is consistent with workers sorting across jobs that trade off outside opportunities and match-specific compensation. That is, a worker may accept low earnings today in exchange for better mobility up the job ladder. A simpler explanation, offered by a reviewer, is that an earnings floor affects low-earnings workers, and is more binding in low-earnings firms. Our empirical model is sufficiently rich to allow for the observed differences in the relationship between earnings and mobility across worker and firm types.

5.2.4 Job Change Event Study

Our diagnostics are related to the event study analysis proposed in Card et al. (2013). They observe that under the exogenous mobility model, there should be a symmetry between earnings increases observed for workers moving from low-earnings firms to high-wage firms and earnings decreases observed for workers moving from high-wage to low-earnings firms. Figure 4 illustrates such an event study applied to our analysis sample. For the analysis, we restrict the sample to workers who are continuously employed over the five year window, change jobs exactly once, and for whom the

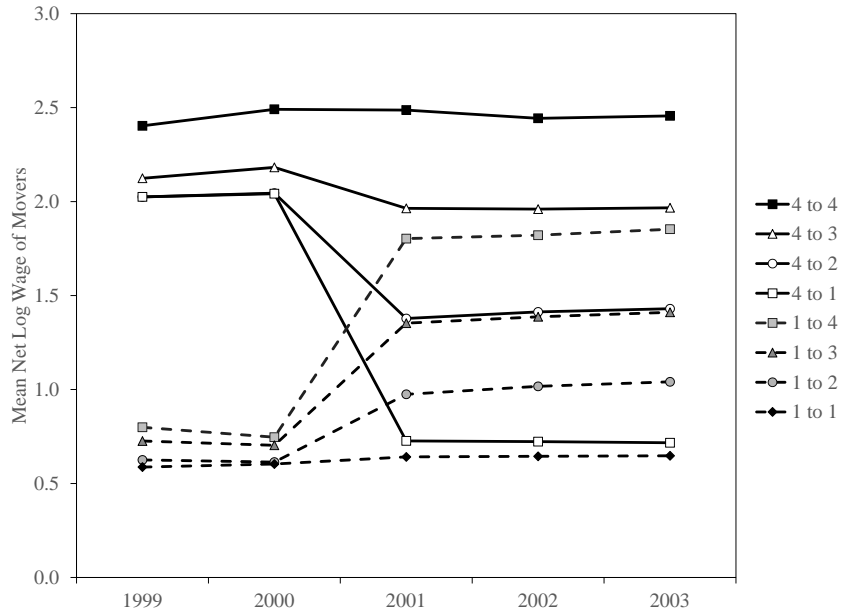


Figure 4: Mean Log Earnings Net of Observed Characteristics for Workers Who Change Jobs in 2001 by Quartile of Firm Effect for Origin and Destination Firms, 1999–2003. See Table OA5 for complete results.

year of job change is 2001. This way we observe workers for two full years before and after the job change. Each job is assigned to its quartile in the firm-effect distribution. Each work history in the sample can be assigned to an origin-quartile / destination-quartile cell. For each year of our data, between 1999–2003, we plot the average log earnings, net of observed characteristics, within each transition cell. If exogenous mobility holds, we expect to see (1) no change in earnings leading up to, or after, the job transition, and (2) that the earnings gains from moving to higher quartiles are symmetric to earnings losses from moving down quartiles. The figure is largely consistent with (1), but less so on (2). Instead, moves from the first to the fourth quartile yield smaller earnings gains than the earnings losses associated with moves from the fourth to the first quartile.

6 Results from Estimation of the Endogenous Mobility Model

We fit the structural model described in Section 3 to the 0.5% LEHD analysis sample using the Gibbs sampler. Our results are based on 9,968 draws, taken in approximately equal proportion, from three parallel runs of the Gibbs sampler. The sampler appears to converge after roughly 500 iterations, but exhibits extensive within-chain autocorrelation. Because of the computational demands, we could not take a large enough sample to eliminate the effects of autocorrelation by thinning the sample by, say, selecting every 1,000th sample to analyze. Instead, we report Monte Carlo standard errors, as described in Appendix Section E.2, that properly account for the serial correlation within each sequence of draws from the sampler.

Before the final estimation, we engaged in a model selection step to determine the number of support points for the distribution of latent person, employer, and match heterogeneity. We base model selection on the criterion that the structural model should explain as much of the variation in earnings as the AKM decomposition. We estimated the model, adding support points to latent heterogeneity distributions until the variance of the structural earnings residual was near the variance of the AKM residual (after removing the orthogonal AKM match effect). That is, we sought a model with granularity sufficient to have as much explanatory power in the earnings equation as the unrestricted AKM model. This approach yields an upper bound on the amount of heterogeneity needed to fit the data, as recommended by Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2013). Details of model selection appear in Appendix E.4.

We report results based on a model in which there are ten points of support for each distribution. That is, there are ten worker types ($L = 10$), ten employer types ($M = 10$), and ten match types ($Q = 10$). A comparison of the correlation with log earnings of ε_{Gibbs} and ε_{AKM} reported in Table 2 shows we were able to obtain a good fit on residual variation. The correlation between the AKM residual and log earnings is $\text{Corr}(y, \varepsilon_{AKM}) = 0.17$, while the correlation between the structural residual and log earnings is $\text{Corr}(y, \varepsilon_{Gibbs}) = 0.24$. In practice, our model may include more support points for the employer heterogeneity than needed. If so, the apparent lack of parsimony does not introduce problems with our results, since we can collapse redundant types in our over-parameterized model.

Our main results use five years of data to limit the influence of drift in the latent heterogeneity

over time. A countervailing concern is that the short window may not provide enough variation to separate the influence of worker, firm, and match heterogeneity. In Appendix F, we report results from estimating our model on the same 0.5% random sample of workers, but where the data are augmented with all available dominant jobs that are recorded in the LEHD infrastructure filesystem between 1990-2010. The results using the extended work histories lead to identical conclusions about the nature of endogenous mobility, and so we focus in the main text on discussing results from the primary analysis sample.

6.1 Summary of Structural Earnings Model Estimates

Figure 5 depicts the posterior distribution of the structural earnings equation parameters using the same scale on the y-axis for each panel. The figures plot the posterior mean $\pm 2 \times MCSE$. The plot showing the posterior mean along with the 5th and 95th percentile of the posterior distribution is nearly identical and appears for reference as Figure OA2 in the online appendix. For consistency with the conventional AKM decomposition, we report the earnings parameters as deviations from the grand mean of log earnings.

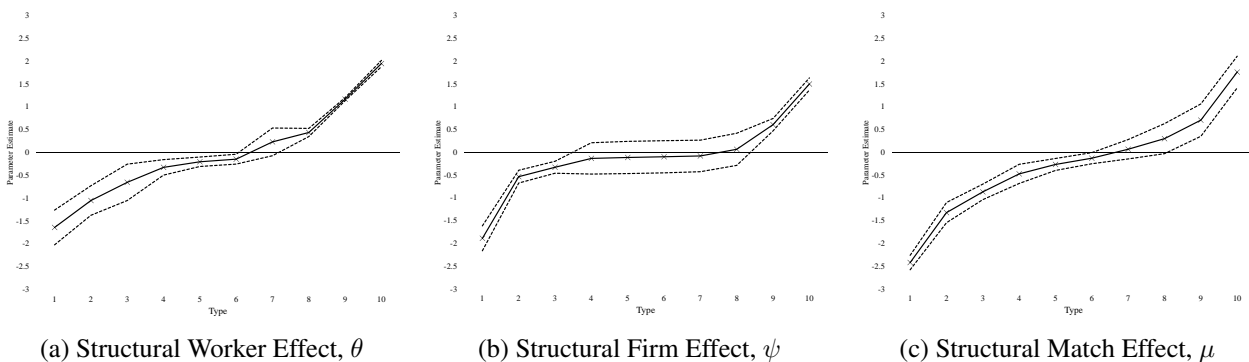


Figure 5: Posterior distribution of earnings equation parameters. Dashed lines indicate the region within $2 \times MCSE$ of the estimate.

Note that there is variation in the estimated earnings parameters on all three heterogeneity dimensions. All of these posterior distributions are extremely tight around the posterior mean of the value of the effect for each type. The dispersion of the match effects is much greater than that of the person effects, θ , or employer effects, ψ . There is very little variation in estimated employer

effects between types 3 and 8. Our model only really detects four or five distinct employer types. Consequently, for the highest-paying firms, those with the largest values of ψ , where two of those four types are located, the confidence intervals around the posterior mean firm effects are very tight.

Table 1: Posterior Distribution of Worker, Firm, and Match Population Heterogeneity

	(1) Worker		(2) Firm			(3) Match	
	Mean	MCSE	Mean	MCSE		Mean	
π_{A1}	0.0828	0.0477	π_{B1}	0.0541	0.0012	π_{K1}	0.0391
π_{A2}	0.0512	0.0179	π_{B2}	0.3374	0.1461	π_{K2}	0.0595
π_{A3}	0.1151	0.0354	π_{B3}	0.0263	0.0341	π_{K3}	0.0814
π_{A4}	0.1305	0.0353	π_{B4}	0.0293	0.0358	π_{K4}	0.1046
π_{A5}	0.0936	0.0230	π_{B5}	0.0273	0.0298	π_{K5}	0.1174
π_{A6}	0.1004	0.0164	π_{B6}	0.0265	0.0264	π_{K6}	0.1229
π_{A7}	0.1376	0.0206	π_{B7}	0.0246	0.0239	π_{K7}	0.0847
π_{A8}	0.1164	0.0094	π_{B8}	0.0083	0.0007	π_{K8}	0.1222
π_{A9}	0.1201	0.0149	π_{B9}	0.4136	0.0254	π_{K9}	0.2028
π_{A10}	0.0523	0.0051	π_{B10}	0.0524	0.0038	π_{K10}	0.0653

Columns (1) and (2) report the posterior distribution of the estimated probability that a worker or firm, respectively, belongs to each latent heterogeneity type. The types are sorted in increasing order by the associated wage heterogeneity component. Column (3) report the marginal distribution across the population of latent match heterogeneity types. There are no associated MCSE estimates because the marginal distribution is computed from the conditional distribution of match quality given the worker and firm types.

Table 1 reports the posterior mean and MCSE of the parameter governing the population distribution of worker types, π_A , the population distribution of employer type, π_B , and the marginal probability for match type, π_K . The latter probability is computed by integrating the conditional probability, $\pi_{k|ab}$ over worker and employer types. All worker types occur in the population with positive probability, however, workers are less likely to be of the highest and lowest type in the population. The case is even more extreme for employers. The distribution of employer types has only four or five distinct points of support. The distribution of employer types is thus very coarse. By contrast, the distribution of match types is the most granular. The marginal distribution of match types is clearly skewed toward higher match quality.

For completeness, Table OA3 in Appendix A reports the posterior mean and MCSE for the parameters associated with observed covariates included in the earnings model. Furthermore, the reader may be concerned that our earnings model, while richly specified, does not explicitly control for serial correlation in the residual. Table OA4 summarizes the posterior distribution of residual autocorrelation at up to three lags, and shows that they are quite small.

Comparison of AKM and Structural Estimates

Table 2 reports correlations, weighted by job duration, among earnings and its components as estimated by least squares (labeled AKM) and from our structural endogenous mobility model (labeled Gibbs). Column (1) reports the correlation of earnings (labeled $\ln w$) with each of the earnings components. In the structural estimates, much more of the variation in earnings is explained by individual heterogeneity than in the AKM estimates. Much less of the variation is explained by employer and match specific heterogeneity. As we will see, though, this is partially due to a strong negative correlation between the structural match effect and the structural firm and worker effects.

Table 2: Correlation Matrix of Wage Equation Parameters: LEHD Data 0.5% Sample, (10, 10, 10) Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	$\ln w$	$X\beta_{AKM}$	θ_{AKM}	ψ_{AKM}	μ_{AKM}	ε_{AKM}	$X\beta_{Gibbs}$	θ_{Gibbs}	ψ_{Gibbs}	μ_{Gibbs}	ε_{Gibbs}
$\ln w$	1.00										
$X\beta_{AKM}$	0.47	1.00									
θ_{AKM}	0.36	-0.57	1.00								
ψ_{AKM}	0.45	0.05	0.10	1.00							
μ_{AKM}	0.20	0.00	0.00	0.00	1.00						
ε_{AKM}	0.17	0.00	0.00	0.00	0.00	1.00					
$X\beta_{Gibbs}$	0.79	0.59	0.13	0.23	0.03	0.02	1.00				
θ_{Gibbs}	0.47	0.13	0.30	0.22	0.00	0.00	0.21	1.00			
ψ_{Gibbs}	0.33	0.08	0.13	0.41	0.08	0.00	0.15	0.01	1.00		
μ_{Gibbs}	0.04	-0.00	0.04	-0.08	0.20	0.00	-0.03	-0.46	-0.46	1.00	
ε_{Gibbs}	0.24	0.01	0.07	0.07	0.12	0.68	0.00	0.00	0.00	0.00	1.00

Table entries are means of the correlation between the indicated variables across 9,968 draws from the Gibbs sampler described in the text.

The AKM estimates of the person and firm effects have a positive correlation of 0.10. This result contrasts somewhat with prior estimates from LEHD data that reported a correlation closer

to zero (Abowd et al. 2003; 2012). This discrepancy may reflect changes over time in the nature of assortative matching, as seems to have been the case in Germany (Card et al. 2013). The AKM match effect, because it is estimated from the least squares residual, is restricted to be uncorrelated with the AKM person and firm effects, as is the case in our subsample. However, the structural person and firm effects are strongly negatively correlated with the structural match effect.

In the structural estimation, the correlation between person and firm effects is much weaker, at 0.01. Note that this need not be the case: our structural model allows for separation and assignment outcomes to be arbitrarily associated with worker and firm types. In the structural model, there is a very strong negative correlation between the structural match effect and the structural worker effect (-0.46) and the structural firm effect (-0.46). The structural estimates also exhibit a positive correlation between observed and unobserved components of individual earnings heterogeneity: $\text{Corr}(X\beta_{Gibbs}, \theta_{Gibbs}) = 0.21$. This result contrasts with the AKM estimates, for which $\text{Corr}(X\beta_{AKM}, \theta_{AKM}) = -0.57$.

The bottom-left panel of Table 2 reports the correlation between the mobility-biased AKM parameters and the structural parameters. This panel provides some insight into the endogenous mobility bias. Observe, first, that there is a positive correlation between the AKM heterogeneity components and their structural counterparts. Second, the AKM estimates of the worker and firm effects combine information from the structural worker, firm, and match effects. In particular, the structural match effect is positively correlated with the AKM person effect and negatively correlated with the AKM firm effect.

These results indicate that the positive correlation between the AKM person and firm effects does not reflect positive assortative matching on latent worker and employer characteristics. The correlation arises instead because high-earnings workers are typically employed on low-paying matches, and the heterogeneity from those matches loads onto the OLS worker and firm effects. The structural model indicates that workers are assigned randomly to firms, and that most earnings variation is associated with worker and match-specific heterogeneity. However, that match-specific heterogeneity acts to compress the earnings distribution. Random assignment is a feature of models of labor market search of the type surveyed in Mortensen (2003) and Pissarides (2000). Our findings are also consistent with the basic framework outlined by Gruetter and Lalive (2009), who

observe that in the presence of endogenous mobility, realized match effects compress the earnings distribution and, as a result, the estimated variance of firm effects in the AKM model will be biased downward relative to the true variance.

Table OA2 in Appendix A reports the same information in the form of a regression of the structural estimates of the earnings decomposition components on the AKM estimates of all components. These regressions compute the conditional expectation of the structure given the AKM estimates. They can be used to compute estimates of the earnings components corrected for endogenous mobility bias from data for which only the AKM estimates are available.

6.2 Summary of Structural Mobility Model

Figures 6 and 7 summarize the mobility model by presenting the stationary distribution of worker-firm pairs and the expected match effect on each such pair. We obtain the steady-state distribution by computing the kernel of the Markov transition matrix implied by our mobility model from the estimated parameters, γ , δ , π_A , π_B and $\pi_{K|AB}$. This gives us, for each worker type, the steady-state probability of observing a worker of that type matched to a particular type firm on a particular type of match.

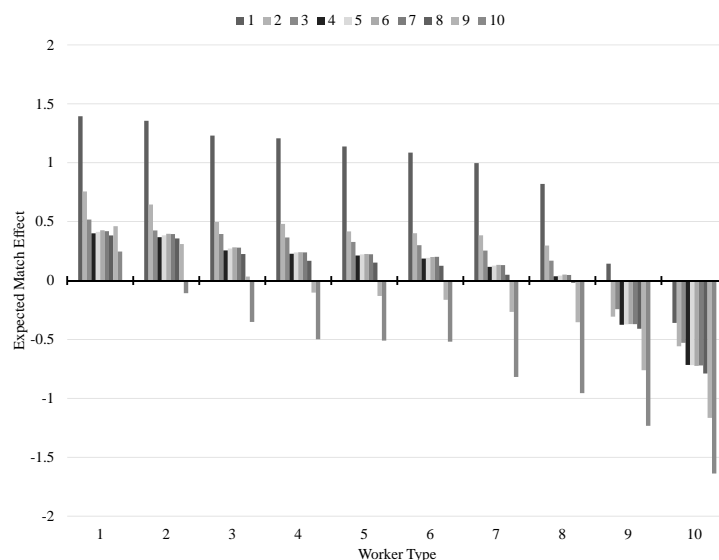


Figure 6: Expected structural match effect by worker/firm type cells. Legend is firm types.

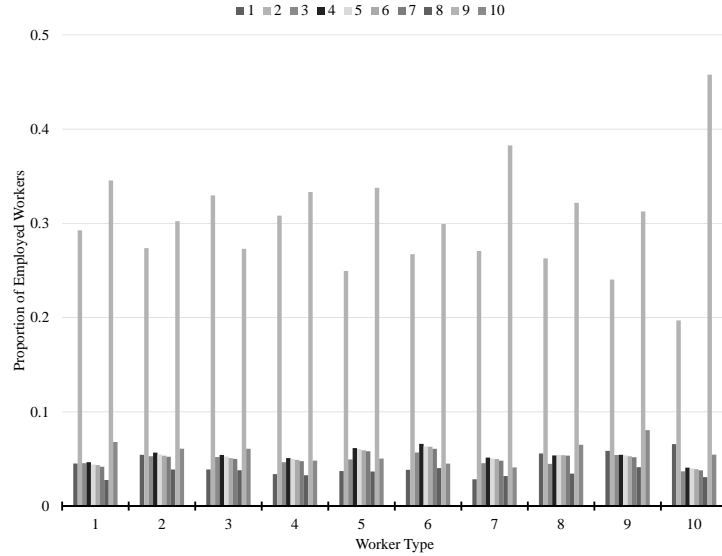


Figure 7: Expected share of matches in steady-state for each worker-type firm-type combination. Legend is firm types.

Figure 6 is structured identically to its AKM analog (Figure 1b). Each bar represents the expected match effect, conditional on a worker type–firm type cell. The bars are grouped by worker type, so that within each worker type, we see the pattern of expected match effects for workers of that type when matched to different firm types. The figure shows two key patterns: the expected match effect is strongly decreasing with firm type and with worker type. Thus, the mobility model in steady-state exhibits the same compression pattern we observed in Table 2, as anticipated in our discussion of the AKM estimates.

Institutions might provide a partial explanation for the observed compression. Workers may have positive estimated match effects when employed in low-earnings firms because those firms are constrained by the minimum wage. This would certainly be a type of match-specific determinant of pay that would depend on worker and firm type, and would also change worker mobility. This could also explain why, when low-type workers move up the firm ladder, they do not get as much of a boost as high-type workers. We thank an anonymous referee for pointing out this interpretation.

Figure 6 shows the expected match effect conditional on observing a given worker-firm combination. Figure 7 gives the probability of observing each worker-firm combination in steady-state. Again, the data are grouped by worker type, so that within each group, the bars show the probabil-

ity of observing a worker in a firm of each type. The pattern for each worker type is very similar to the population distribution of firm types reported in Table 1. There is some evidence of a selection effect: workers of type 10 are roughly 50 percent more likely to be observed in high-wage firms than workers of type 1. This is consistent with the weak evidence of positive assortative matching from Table 2. Overall, it appears that worker-firm matches are sampled almost randomly in steady-state.

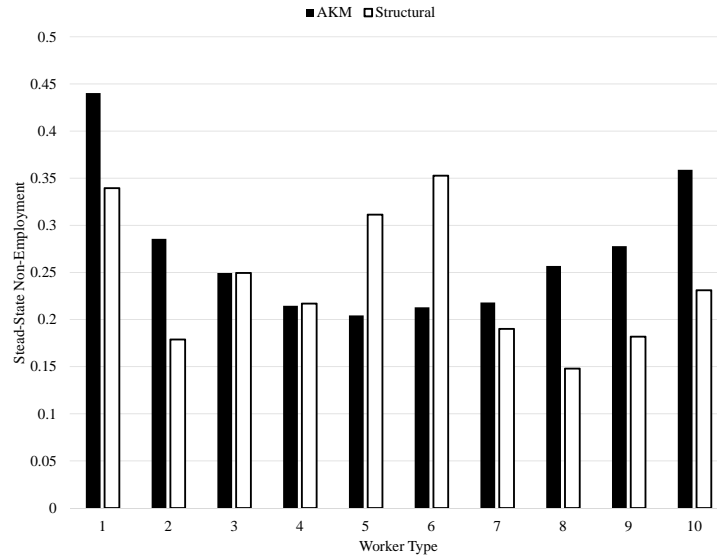


Figure 8: Probability of not being employed in the sample frame by worker effect decile (AKM) and latent worker type (Structural)

Figure 8 shows the probability of non-employment, conditional on the latent worker type. For comparison, we also show the corresponding probabilities when workers are grouped by the decile of their AKM worker effect. The estimates from the structural model show the non-employment probability is weakly decreasing in the latent worker type. That is, more highly paid workers are more likely to be employed. This standard mover-stayer result is not assumed in estimation. By contrast, the results based on the AKM worker effect deciles indicate that highly-paid and low-paid workers are less likely to be employed than workers in the middle of the distribution. The non-employment probabilities may seem high because they incorporate transitions out of the labor force and into employment outside our three-state sample frame in addition to unemployment.

Finally, Figure 9 displays the structural analog to Figure 3. It reports, for workers who leave

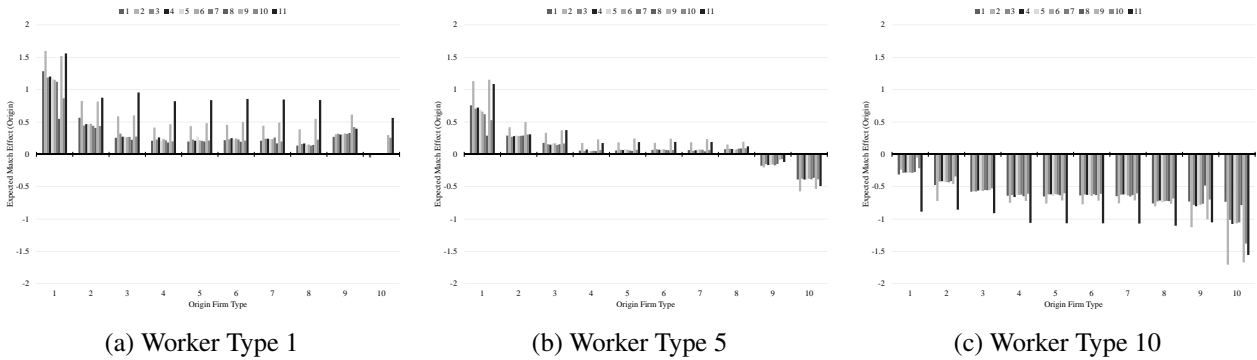


Figure 9: Expected match effect within origin/destination employer type cells, disaggregated by worker type. Legend is firm types.

their job, the expected match effect on the origin job conditional on the types of the origin and destination job. The plots are disaggregated by worker type. The plots illustrate the compression pattern that emerged in the discussion of Table 2. The match effect is negatively correlated with the origin firm type, as well as with worker type. Furthermore, conditional on worker type and origin-firm type, the magnitude of the match effect does not appear to provide nearly as much power to predict the destination job type as was the case in Figure 3 based on the mobility-biased AKM estimates.

6.3 Validation: Relationship with Revenue

If the AKM estimates are biased by endogenous mobility, we should observe a difference in the relationship between employer revenues and the earnings heterogeneity components estimated from our structural model. Table 3 reports the results of estimating a firm-level regression of log revenue per worker on the estimated firm effect, the average worker effect within firm, and the average match effect within firm. For this analysis, we restrict the sample to jobs from 2002, which is the reference year for the revenue data in the 2002 Economic Census. After this restriction, the sample consists of 60,589 firms. The columns under (1) report the parameter estimates and standard errors from the regression onto firm-level aggregates of the structural earnings components. The columns under (2) reports the estimates from the regression onto firm-level aggregates of the uncorrected AKM earnings components.

Table 3: Regression of Log Revenue Per Worker on Structural and AKM Estimates of Wage Decomposition Components

	(1)		(2)	
	Structural		AKM	
	Coef	Ste	Coef	Ste
Firm Avg. θ	0.1967	0.0129	0.0637	0.0062
ψ	0.2349	0.0133	0.510	0.0140
Firm Avg. μ	0.1732	0.0141	0.007	0.0194
Firm Avg. $X\beta$	0.0398	0.0077	0.0354	0.0059
Intercept	3.3526	0.0687	3.4542	0.0445
N	60, 589		60, 589	

Coefficient estimates from firm-level regressions of log revenue per worker onto firm-level averages of estimated earnings heterogeneity. Panel (1) uses the estimated wage components corrected for endogenous mobility bias from the structural model. Panel (2) uses the uncorrected estimates computed using the AKM decomposition under the assumption of exogenous mobility. The sample is based on data from 2002, the reference year for revenue data in the 2002 Economic Census. The revenue data are multiply imputed where missing, and the reported coefficients and standard errors include Rubin’s correction for imputation uncertainty.

When we use the AKM components of earnings heterogeneity, revenue per worker is strongly correlated with the AKM firm effect, and only negligibly correlated with the average worker and match effects. When, by contrast, we use the estimates of earnings heterogeneity from the endogenous mobility model, the average match effect and average worker effect are much more strongly correlated with revenue. We also find that the correlation between the structural firm effect and log revenue per worker is considerably smaller than when the AKM firm effect is used: 0.2349 versus 0.5100.

We interpret these results as supporting our correction for endogenous mobility. Revenue per worker should be a function of total human capital, both in the form of average worker quality and match quality, and any productive advantage accorded by the firm-specific wage premium. In the uncorrected estimates, there is no relationship between worker quality (average worker effect) or match quality and revenue. The variation in revenue per worker, which drives workers across jobs, loads entirely onto the firm effect. After the correction for endogenous mobility, worker heterogeneity and match quality have the relationships we would expect with revenue per worker. It remains the case that more productive firms – those with greater revenue per worker – are also high-

wage firms. Finally, we note that the regression of firm revenue on AKM earnings components is potentially biased by sampling errors in the component estimates. Under some conditions, this problem might not affect the regression onto Gibbs posteriors.

7 Conclusion

Our analysis confirms the importance of relaxing the assumption that job mobility is exogenous. Exogenous mobility is essential to any direct causal interpretation of the AKM earnings decomposition as well as for downstream applications that use the estimated worker and firm effects as measures of skill and compensation policy. Our findings indicate that the assumption of exogenous mobility is rejected in data from the LEHD program. Furthermore, the analysis of residuals from the AKM decomposition indicates the presence of omitted match-specific heterogeneity. That omitted match-specific heterogeneity is predictive of the type of firm to which a worker moves.

To relax the exogenous mobility assumption, we estimate a latent-type model that incorporates these features. We allow for a match effect that is arbitrarily correlated with worker and firm heterogeneity, and we allow the match effect to drive both the decision to separate and the type of firm to which the worker moves. Our results indicate that allowing for correlated match effects has a strong effect on estimated worker- and firm-specific heterogeneity. Validation against firm revenue data suggest that our corrections for endogeneity move the estimated effects in an economically meaningful direction.

Our analysis is subject to some caveats. Our model is extremely computationally intensive, requiring us to estimate on a sample of the LEHD data. While we have used a dense sub-sample, the implications of sampling in relational, or network data, particularly for this sort of analysis remain poorly understood. The parallelization afforded by the graph coloring algorithm can be scaled up, but not within the computing facilities available through the Census Bureau. Additionally, we have considered one model that relaxes the exogenous mobility assumption. The model we consider is consistent with the residual diagnostics, but other models are possible.

For researchers working with these models, our results indicate that it is important to test for failure of the exogenous mobility assumption. When it fails, it may also be important to attempt

to correct for endogenous mobility bias. We find that the OLS estimates from the AKM model are positively correlated with conceptually appropriate effects that were estimated correcting for endogenous mobility. However, correcting for endogenous mobility has a substantial affect on the relationship between worker and firm earnings heterogeneity, and on the relationship of these components with firm revenue per worker. These findings recommend that caution is warranted when interpreting worker and firm effects estimated under the AKM assumptions, as originally noted by Abowd and Kramarz (1999).

Acknowledgements

We have benefitted from discussion with Joseph Altonji, David Card, John Eltinge, Patrick Kline, Francis Kramarz, Jesse Rothstein, Richard Upward, and Lars Vilhuber, along with participants at numerous seminars. Nellie Zhao provided expert research assistance. The research in this paper was previously circulated in two working papers: “Endogenous Mobility,” and “How Important is Endogenous Mobility for Measuring Employer and Employee Heterogeneity.” Abowd acknowledges direct support from NSF Grants SES-0339191, CNS-0627680, SES-0922005, TC-1012593, and SES-1131848. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This research uses data from the Census Bureau’s Longitudinal Employer-Household Dynamics Program, which was partially supported by National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

Supplementary Materials

Abowd et al. (2017) [DOI: [10.5281/zenodo.376600](https://doi.org/10.5281/zenodo.376600)] provide an archive of SAS and MATLAB code to build the analysis data, estimate the AKM and endogenous mobility models, calculate test statistics, and generate statistical summaries described in the article. The archive also includes the output files released by the Census Bureau after disclosure avoidance review.

Bibliography

- Abowd, J. M., Creecy, R. H. and Kramarz, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data, *Technical Report TP-2002-06*, LEHD, U.S. Census Bureau.
- Abowd, J. M. and Kramarz, F. (1999). The analysis of labor markets using matched employer-employee data, in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3 of *Handbook of Labor Economics*, Elsevier, chapter 40, pp. 2629–2710.
- Abowd, J. M., Kramarz, F., Lengerman, P., McKinney, K. and Roux, S. (2012). Persistent inter-industry wage differences: Rent-sharing and opportunity costs, *IZA Journal of Labor Economics* **1**(1): 1–7.
- Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). High wage workers and high wage firms, *Econometrica* **67**(2): 251–333.
- Abowd, J. M., Kramarz, F., Pérez-Duarte, S. and Schmutte, I. M. (2015). Sorting between and within industries: A testable model of assortative matching. Labor Dynamics Institute Working Paper No. 21.
- Abowd, J. M., Lengermann, P. and McKinney, K. L. (2003). The measurement of human capital in the U.S. economy, *Technical Report TP-2002-09*, LEHD, U.S. Census Bureau.
- Abowd, J. M., McKinney, K. L. and Schmutte, I. M. (2017). Replication Archive for "Modeling Endogenous Mobility in Earnings Determination".
URL: <https://doi.org/10.5281/zenodo.376600>
- Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M. and Woodcock, S. (2009). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators, in T. Dunne, J. Jensen and M. Roberts (eds), *Producer Dynamics: New Evidence from Micro Data*, Chicago: University of Chicago Press for the National Bureau of Economic Research, pp. 149–230.

- Andrews, M. J., Gill, L., Schank, T. and Upward, R. (2008). High wage workers and low wage firms: negative assortative matching or limited mobility bias?, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(3): 673–697.
- Bonhomme, S., Lamadon, T. and Manresa, E. (2015). A distributional framework for matched employer employee data, *Unpublished manuscript*, University of Chicago.
- Card, D., Heining, J. and Kline, P. (2013). Workplace heterogeneity and the rise of West German wage inequality, *The Quarterly Journal of Economics* **128**(3): 967–1015.
- Combes, P.-P., Duranton, G. and Gobillon, L. (2008). Spatial wage disparities: Sorting matters!, *Journal of Urban Economics* **63**: 723–742.
- Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(2): 363–375.
- Eeckhout, J. and Kircher, P. (2011). Identifying sorting—in theory, *Review of Economic Studies* **78**: 872–906.
- Gebremedhin, A. H., Manne, F. and Pothén, A. (2005). What color is your Jacobian? Graph coloring for computing derivatives, *SIAM Rev.* **47**(4): 629–705.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
- Gelman, A., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models, *Statistics and Computing* **24**(6): 997–1016.
- Geyer, C. (2011). Introduction to Markov Chain Monte Carlo, *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, Chapman and Hall/CRC.
- Geyer, C. J. (1992). Practical Markov Chain Monte Carlo, *Statistical Science* **7**(4): pp. 473–483.
- Gibbons, R., Katz, L. F., Lemieux, T. and Parent, D. (2005). Comparative advantage, learning and sectoral wage determination, *Journal of Labor Economics* **23**: 681–723.

- Gruetter, M. and Lalive, R. (2009). The importance of firms in wage determination, *Labour Economics* **16**: 149–160.
- Helwege, J. (1992). Sectoral shifts and interindustry wage differentials, *Journal of Labor Economics* **10**(1): 55–84.
- Hoff, P. D., Raftery, A. and Hancock, M. (2002). Latent space approaches to social network analysis, *Journal of the American Statistical Association* **97**(460): 1090–1098.
- Iranzo, S., Schivardi, F. and Tosetti, E. (2008). Skill dispersion and firm productivity: An analysis with matched employer-employee data, *Journal of Labor Economics* **26**: 247–285.
- Koedel, C. and Betts, J. R. (2010). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique, *Education Finance and Policy* **5**: 54–81.
- Kosorok, M. R. (2000). Monte carlo error estimation for multivariate markov chains, *Statistics & Probability Letters* **46**(1): 85 – 93.
- Kramarz, F., Machin, S. and Ouazad, A. (2014). Using compulsory mobility to identify school quality and peer effects, *Oxford Bulletin of Economics and Statistics* .
- Krishna, P., Poole, J. P. and Senses, M. Z. (2014). Wage effects of trade reform with endogenous worker mobility, *Journal of International Economics* **93**(2): 239 – 252.
- Mortensen, D. (2003). *Wage Dispersion: Why Are Similar Workers Paid Differently?*, Zeuthen lecture book series, MIT Press, Cambridge, Mass.
- Newman, M. E. J. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks, *Proceedings of the National Academy of Sciences of the United States* **104**: 9564–9569.
- Pissarides, C. A. (2000). *Equilibrium Unemployment Theory*, 2nd ed edn, MIT Press, Cambridge, Mass.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay and student achievement, *Quarterly Journal of Economics* **125**(1): 175–214.

- Schmutte, I. M. (2014). Free to move? A network analytic approach for learning the limits to job mobility, *Labour Economics* **29**: 49 – 61.
- Schmutte, I. M. (2015). Job referral networks and the determination of earnings in local labor markets, *Journal of Labor Economics* **33**(1): 1–32.
- Shimer, R. (2005). The assignment of workers to jobs in an economy with coordination frictions, *Journal of Political Economy* **113**: 996.
- Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statistics, 3rd edn, Springer-Verlag.
- U.S. Census Bureau (2006). Comparative statistics: 2002, *Core Business Statistics Series EC02-00C-COMP*, U.S. Census Bureau, 2002 Economic Census.
- Watanabe, S. (2013). A widely applicable Bayesian information criterion, *Journal of Machine Learning Research* **14**(1): 867–897.
- Woodcock, S. (2008). Wage differentials in the presence of unobserved worker, firm, and match heterogeneity, *Labour Economics* **15**: 771–793.

Appendices

A Online Appendix: Supplemental Tables and Figures

Table OA1: Covariance Matrix of Earnings Equation Parameters: LEHD Data 0.5% Sample, (10, 10, 10) Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	$\ln w$	$X\beta_{AKM}$	θ_{AKM}	ψ_{AKM}	μ_{AKM}	ε_{AKM}	$X\beta_{Gibbs}$	θ_{Gibbs}	ψ_{Gibbs}	μ_{Gibbs}	ε_{Gibbs}
$\ln w$	1.949	0.920	0.629	0.278	0.078	0.056	0.919	0.551	0.327	0.045	0.107
$X\beta_{AKM}$	0.920	1.989	-1.022	0.029	-0.001	0.001	0.693	0.151	0.075	-0.005	0.006
θ_{AKM}	0.629	-1.022	1.599	0.056	0.000	0.000	0.135	0.315	0.112	0.038	0.029
ψ_{AKM}	0.278	0.029	0.056	0.194	0.000	0.000	0.085	0.087	0.126	-0.029	0.010
μ_{AKM}	0.078	-0.001	0.000	0.000	0.080	0.000	0.007	0.000	0.016	0.044	0.011
ε_{AKM}	0.056	0.001	0.000	0.000	0.000	0.056	0.004	0.000	0.000	0.000	0.052
$X\beta_{Gibbs}$	0.919	0.693	0.135	0.085	0.007	0.004	0.701	0.146	0.090	-0.017	0.000
θ_{Gibbs}	0.551	0.151	0.315	0.087	0.000	0.000	0.146	0.707	0.003	-0.305	0.000
ψ_{Gibbs}	0.327	0.075	0.112	0.126	0.016	0.000	0.090	0.003	0.485	-0.251	0.000
μ_{Gibbs}	0.045	-0.005	0.038	-0.029	0.044	0.000	-0.017	-0.305	-0.251	0.619	0.000
ε_{Gibbs}	0.107	0.006	0.029	0.010	0.011	0.052	0.000	0.000	0.000	0.000	0.107

Table entries are means of the covariance between the indicated variables across 9,968 draws from the Gibbs sampler described in the text using the estimation sample, with number of person-year observations 395,930.

Table OA2: Regression of Structural Earnings Decomposition Components on AKM Estimates of Earnings Decomposition Components

	$X\beta_{Gibbs}$	θ_{Gibbs}	ψ_{Gibbs}	μ_{Gibbs}	ε_{Gibbs}
$X\beta_{AKM}$	0.2600 (.0010)	0.0402 (.0031)	0.0129 (.0017)	0.2791 (.0113)	0.0159 (.0004)
θ_{AKM}	0.2389 (.0021)	0.2094 (.0105)	0.0559 (.0022)	0.2116 (.0071)	0.0265 (.0008)
ψ_{AKM}	0.3305 (.0012)	0.3812 (.0475)	0.6308 (.0438)	-0.2547 (.0181)	0.0413 (.0020)
μ_{AKM}	0.0944 (.0005)	0.0006 (.0001)	0.2012 (.0114)	0.5519 (.0137)	0.1374 (.0028)
ε_{AKM}	0.0637 (.0003)	-0.0007 (.0001)	-0.0002 (.0000)	-0.0048 (.0002)	0.9359 (.0003)
Constant	7.2656 (.0574)	-0.2421 (.0186)	-0.0344 (.0015)	-2.2883 (.0774)	-0.1258 (.0035)

Results from running a regression of the earnings components estimated under the endogenous mobility model on earnings components estimated using the AKM decomposition. The reported values are the mean parameter estimate and the correlated-draw Monte Carlo standard errors across 9,968 draws from the Gibbs sampler using the estimation sample, with number of person-year observations 395,930.

Table OA3: Observed Covariate Parameters: AKM and Structural Estimates

Variable	(1)	(2)	
	AKM Model Estimate	Structural Model Mean	(MCSE)
<i>age</i>	0.5555	0.6321	(.0064)
<i>age</i> ²	-0.1864	-0.2087	(.0023)
<i>age</i> ³	0.0277	0.0311	(.0003)
<i>age</i> ⁴	-0.0016	-0.0018	(.0000)
<i>female</i> × <i>age</i>	-0.2534	0.0031	(.0021)
<i>age</i> ²	0.0866	-0.0091	(.0010)
<i>age</i> ³	-0.0136	0.0019	(.0002)
<i>age</i> ⁴	0.0008	-0.0001	(.0000)
<i>black</i> × <i>age</i>	-0.2226	0.0157	(.0025)
<i>age</i> ²	0.0788	-0.0160	(.0019)
<i>age</i> ³	-0.0127	0.0035	(.0004)
<i>age</i> ⁴	0.0008	-0.0002	(.0000)
<i>hispanic</i> × <i>age</i>	0.0644	0.0369	(.0063)
<i>age</i> ²	-0.0323	-0.0173	(.0050)
<i>age</i> ³	0.0065	0.0042	(.0012)
<i>age</i> ⁴	-0.0004	-0.0002	(.0001)
<i>sixq2</i>	0.8851	0.7553	(.0069)
<i>sixq3</i>	1.6680	1.6226	(.0088)
<i>sixq4</i>	2.0301	2.1874	(.0115)
<i>sixq5</i>	2.3843	2.6303	(.0120)
<i>sixq6</i>	2.6301	2.7496	(.0103)
<i>sixqleft</i>	-0.1161	-0.0629	(.0037)
<i>sixqright</i>	-0.1054	-0.0301	(.0042)
<i>sixq4th</i>	-0.3500	0.0882	(.0009)
<i>sixqinter</i>	0.2688	-0.3877	(.0030)
<i>yr2000</i>	0.0233	0.0044	(.0007)
<i>yr2001</i>	0.0308	-0.0020	(.0008)
<i>yr2002</i>	0.0494	-0.0089	(.0012)
<i>yr2003</i>	0.0611	-0.0097	(.0014)
σ		0.3277	(.0039)
Num. Obs.	60,123,894	395,930	

Table entries are parameters on time-varying characteristics included in both the AKM model and the structural endogenous mobility model. Column (1) reports parameter estimates from the fit of the AKM model to the LEHD analysis population. Columns under (2) report the posterior means and Monte Carlo Standard Errors for parameters on the indicated control variable based on 9,968 draws from the Gibbs sampler.

Table OA4: Serial correlation in structural residuals

	Correlation Coeff.	MCSE
$\rho_{t,t-1}$	0.1153	0.0054
$\rho_{t,t-2}$	-0.0294	0.0060
$\rho_{t,t-3}$	-0.0805	0.0067

Each row reports the posterior mean and MCSE across 9,968 draws from the Gibbs sampler of the within-worker correlation in the unexplained residual portion of earnings from the endogenous mobility model.

Table OA5: Mean Log Earnings Net of Observed Characteristics for Workers Who Change Jobs in 2001 by Quartile of Firm Effect for Origin and Destination Firms, 1999–2003

Transition Cell	1999	2000	2001	2002	2003
1 to 1	0.587	0.604	0.642	0.645	0.647
1 to 2	0.625	0.614	0.975	1.017	1.041
1 to 3	0.726	0.703	1.353	1.387	1.411
1 to 4	0.799	0.746	1.803	1.821	1.853
2 to 1	1.017	1.030	0.756	0.765	0.759
2 to 2	1.208	1.239	1.264	1.264	1.273
2 to 3	1.354	1.356	1.585	1.602	1.622
2 to 4	1.496	1.488	1.996	2.013	2.041
3 to 1	1.449	1.437	0.741	0.770	0.760
3 to 2	1.567	1.577	1.367	1.394	1.398
3 to 3	1.806	1.834	1.857	1.845	1.860
3 to 4	1.997	2.004	2.205	2.205	2.236
4 to 1	2.025	2.042	0.727	0.723	0.717
4 to 2	2.025	2.045	1.378	1.413	1.430
4 to 3	2.124	2.182	1.964	1.960	1.967
4 to 4	2.403	2.491	2.487	2.443	2.456

The table entries are means of log earnings net of the effect of observed time-varying characteristics for a specific year and transition cell. The sample is the LEHD analysis population described in Section 5.1.1 who change jobs exactly once between 1999 and 2003, and where the year of job transition is 2001. This sample follows 566,300 workers across 183,100 unique firms for a total of 2,832,000 person-year observations. Each job is assigned to a quartile based on the estimated AKM firm effect. The “Transition Cell” column indicates the quartile of the origin and destination job. Figure 4 displays a selection of these transition summaries.

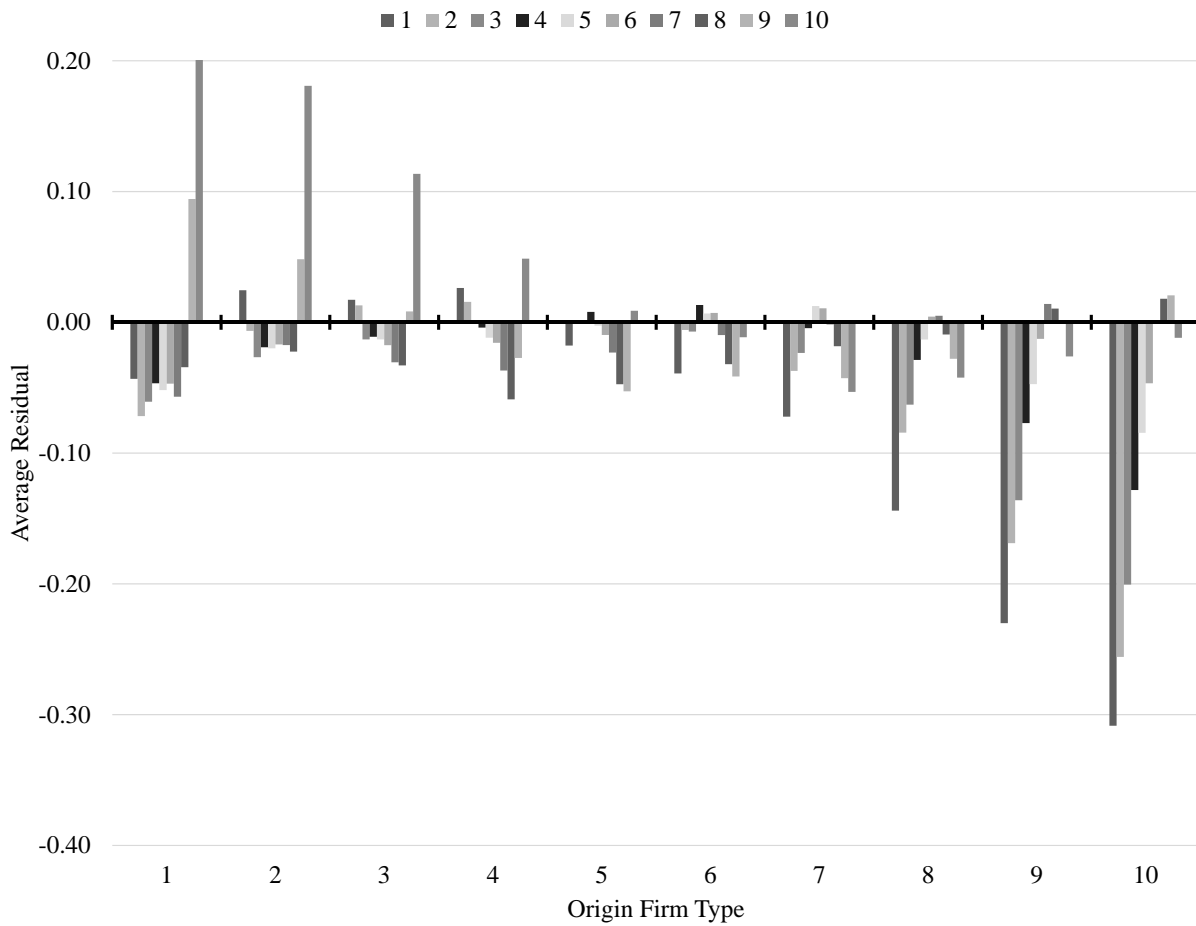
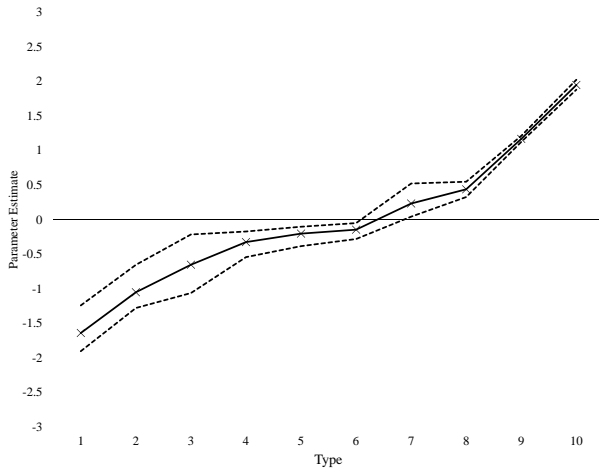
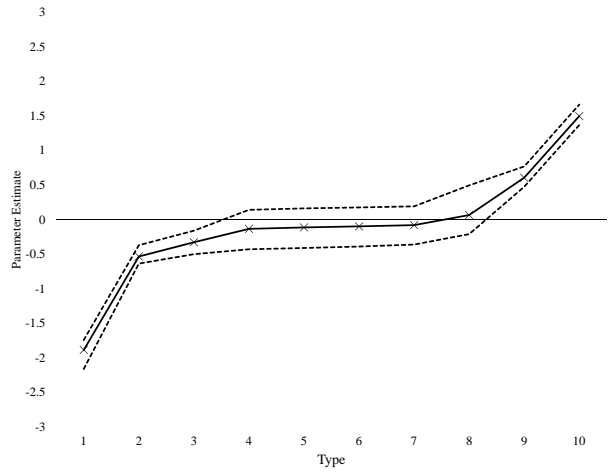


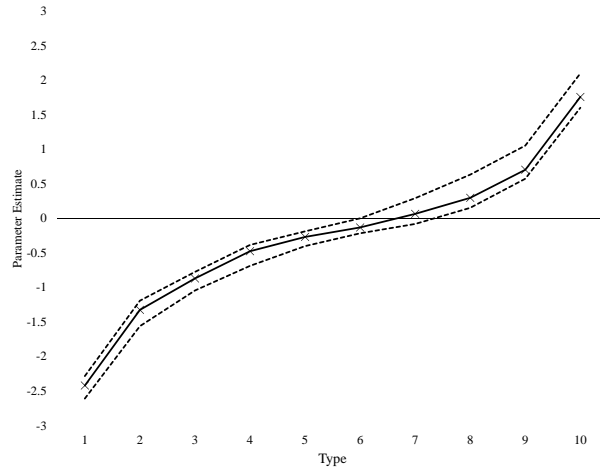
Figure OA1: Mean change in the AKM residual within origin/destination firm effect decile. Legend is for destination firm types.



(a) Structural Worker Effect, θ



(b) Structural Firm Effect, ψ



(c) Structural Match Effect, μ

Figure OA2: Posterior distribution of wage equation parameters. The solid line indicates the posterior mean. Dashed lines indicate the 5th and 95th percentiles.

B Online Appendix: Formal Test of Endogenous Mobility

To implement the tests, we discretize estimated person effects, firm effects, and residuals onto a fixed support. The quantiles that define the support points are calculated from a point-in-time snapshot of the distribution of dominant jobs in progress as of April 1, 2002. That distribution is restricted to full-time, full-year jobs held by individuals age 18-70. Finally, in testing, we use all 465 million dominant job observations for workers 18-70 that occur between 1999 and 2002. Test 1, the match effects test, uses data for about 104 million job changers during 1999-2004, inclusive. Test 2, the productive workforce test, uses data for about 4 million firms alive in 2001.

B.1 Data Preparation and Definitions

Given the fitted values from the AKM decomposition, we select the sample of individuals and employers active at the beginning of 2002, quarter 2 (April 1, 2002). For this sample, we compute deciles from the estimated $\hat{\theta}_i$, $\hat{\psi}_{J(i,t)}$, and $\hat{\varepsilon}_{it}$ as described above. Using the estimated deciles, we discretize each component of the decomposition onto 10 fixed points of support. We adopt the following notation:

$$Q(z) = a \text{ denotes quantile } a \text{ for } z \in \{\theta, \psi, \varepsilon\}$$

and

$$\#Q(z) \text{ denotes then number of quantiles for } z \in \{\theta, \psi, \varepsilon\}.$$

In the tests presented below, we use deciles, so $\#Q(z) = 10$.

B.2 Test Statistic 1: Match Effects Test

Under the hypothesis of exogenous mobility, the match effect for a given individual–employer pair can be estimated using the average residual for the most recent completed job at j by i . We denote these match effects as $\bar{\varepsilon}_{it-1}$ for those individuals who change employers between periods $t-1$ and t . Formally,

$$\bar{\varepsilon}_{it-1} = \frac{\sum_{\{s|J(i,s)=j \wedge s < t \wedge J(i,s) \neq J(i,t)\}} \hat{\varepsilon}_{is}}{\sum 1_{\{s|J(i,s)=j \wedge s < t \wedge J(i,s) \neq J(i,t)\}}}$$

An individual for whom $\bar{\varepsilon}_{it-1} > 0$ received wage payments while employed at $J(i, t-1) = j$ that exceeded their expected value, again under the hypothesis of exogenous mobility. The opposite is true for individuals for whom $\bar{\varepsilon}_{it-1} < 0$.

B.2.1 Derivation of the Match Effects Test Statistic

To form a test statistic that captures the potential for $\bar{\varepsilon}_{it-1}$ to be predictive of the next employer type, we count all (i, t) pairs where $J(i, t-1) \neq J(i, t)$ (job changers) in quantiles of the components

$\hat{\theta}_i, \hat{\psi}_{J(i,t-1)}, \hat{\psi}_{J(i,t)}$, and $\bar{\varepsilon}_{it-1}$:

$$n_{abcd} = \sum_{\{i,t|J(i,t-1) \neq J(i,t)\}} 1 \left\{ \begin{array}{l} Q(\hat{\theta}_i) = a \wedge \\ Q(\hat{\psi}_{J(i,t-1)}) = b \wedge \\ Q(\hat{\psi}_{J(i,t)}) = c \wedge \\ Q(\bar{\varepsilon}_{it-1}) = d \end{array} \right\}. \quad (\text{B-1})$$

The joint probability of observing n_{abcd} is

$$\pi_{abcd} = \Pr \{ Q(\theta_i) = a \wedge Q(\psi_{J(i,t-1)}) = b \wedge Q(\psi_{J(i,t)}) = c \wedge Q(\bar{\varepsilon}_{it-1}) = d \}.$$

Exogenous mobility implies that the match effect from period $t - 1$ should not be predictive of the transition from $\psi_{J(i,t-1)}$ to $\psi_{J(i,t)}$ for an individual with θ_i . This hypothesis can be formalized as conditional independence of the outcome

$$(Q(\theta_i) = a \wedge Q(\psi_{J(i,t-1)}) = b \wedge Q(\psi_{J(i,t)}) = c)$$

from $Q(\bar{\varepsilon}_{it-1}) = d$. In terms of the joint probabilities we compute

$$X_{\nu_1}^2 = \text{Test}(\pi_{abcd} = \pi_{abc+} \pi_{+++d}) \quad (\text{B-2})$$

where the subscript $+$ denotes the marginal distribution with respect to the indicated dimension, and degrees of freedom are given by

$$\nu_1 = (\#(Q(\theta_i)) \times \#Q(\psi_{J(i,t-1)}) \times \#Q(\psi_{J(i,t)}) - 1) \times (\#Q(\bar{\varepsilon}_{it-1}) - 1).$$

B.2.2 Computation of the Match Effects Test

We compute the test statistic (B-2) by direct calculation of the chi-squared statistic from the 4-way contingency table defined by the discretized earnings heterogeneity under the conditional independence assumption $\pi_{abcd} = \pi_{abc+} \pi_{+++d}$. The population of job changers consists of individuals i for whom $J(i, t - 1) \neq J(i, t)$ for $t = 1999, \dots, 2003$. The entire population of individuals and employers was used to compute the quantiles of the $\hat{\theta}_i, \hat{\psi}_{J(i,t-1)}, \hat{\psi}_{J(i,t)}$, and $\bar{\varepsilon}_{it-1}$ distributions. Then the counts (B-1) were tabulated using all observations in the job-changer population and used to compute the relevant marginal frequencies for the test.

B.3 Test Statistic 2: Productive Workforce Test

Our second test considers the implications of exogenous mobility for the employer's choice of workforce distributions over θ_i . The average amount by which wages deviate from their expectations, under exogenous mobility, for a given workforce at a point in time can be computed as the average residual for all employees at $J(i, t) = j$ in year t

$$\tilde{\varepsilon}_{jt} = \frac{\sum_{\{i|J(i,t)=j\}} \hat{\varepsilon}_{it}}{\sum 1 \{i|J(i,t) = j\}}.$$

An employer for whom $\tilde{\varepsilon}_{jt} > 0$ has paid higher than expected wages in period t ; and the opposite is true for $\tilde{\varepsilon}_{jt} < 0$. Although there could be many reasons for this, we will refer to $\tilde{\varepsilon}_{jt}$ as a measure of workforce productivity. However, the exogenous mobility hypothesis is silent about the meaning of $\tilde{\varepsilon}_{jt}$. What matters is its relationship to the within-employer distribution of θ_i . If $\tilde{\varepsilon}_{js}$ is predictive of the within-employer distribution of θ_i for some period $t > s$, given ψ_j , then exogenous mobility fails because the distribution of future employment depends on residuals in the theoretical AKM decomposition.

To implement this test, consider two periods $s < t$ and all employers with strictly positive employment in period s . Compute the counts

$$n_{abc|s} = \sum_j \left\{ 1 \{Q(\psi_j) = a \wedge Q(\tilde{\varepsilon}_{js}) = c\} \times \sum_{\{i|J(i,s)=j \wedge Q(\psi_j)=a\}} Q(\theta_i) = b \right\}$$

and

$$n_{abc|t} = \sum_j \left\{ 1 \{Q(\psi_j) = a \wedge Q(\tilde{\varepsilon}_{js}) = c\} \times \sum_{\{i|J(i,t)=j \wedge Q(\psi_j)=a\}} Q(\theta_i) = b \right\}.$$

Note that the two counts are not independent because they condition on the same distribution of employers alive in period s . Let

$$\pi_{abc|s} = \Pr \{Q(\psi_j) = a \wedge (Q(\theta_i) = b|s) \wedge Q(\tilde{\varepsilon}_{js}) = c\}$$

and

$$\pi_{abc|t} = \Pr \{Q(\psi_j) = a \wedge (Q(\theta_i) = b|t) \wedge Q(\tilde{\varepsilon}_{js}) = c\}.$$

Then, the statistic for testing the conditional independence of the within-employer distribution over θ_i with respect to the residual is

$$X_{\nu_2}^2 = Test \left(\ln \left(\frac{\pi_{abc|s}}{\pi_{abc|t}} \right) = \ln \left(\frac{\pi_{ab+|s}}{\pi_{ab+|t}} \right) \right) \quad (\text{B-3})$$

with degrees of freedom $\nu_2 = (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1) + (\#Q(\psi_j) - 1) \times (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1)$.

B.3.1 Derivation of the Productive Workforce Test Statistic

To see why the test in equation (B-3) is correct, consider the log-linear model

$$\begin{aligned} \ln \left(\frac{\pi_{abc|s}}{\pi_{abc|t}} \right) &= (\mu_{a|s} - \mu_{a|t}) + (\mu_{b|s} - \mu_{b|t}) + (\mu_{c|s} - \mu_{c|t}) \\ &\quad + (\gamma_{ab|s} - \gamma_{ab|t}) + (\gamma_{ac|s} - \gamma_{ac|t}) + (\gamma_{bc|s} - \gamma_{bc|t}) \\ &\quad + (\rho_{abc|s} - \rho_{abc|t}) \end{aligned}$$

where the notation is as follows:

- $\mu_{z|t}$ denotes main effects of $z \in \{Q(\psi_j), Q(\theta_i), Q(\tilde{\varepsilon}_{js})\}$ in period t ,
- $\gamma_{yz|t}$ denotes 2-way interactions of $(y, z) \in \{Q(\psi_j), Q(\theta_i), Q(\tilde{\varepsilon}_{js})\}$ in period t ,
- $\rho_{xyz|t}$ denotes 3-way interactions of $(x, y, z) \in \{Q(\psi_j), Q(\theta_i), Q(\tilde{\varepsilon}_{js})\}$ in period t .

The change in main effects of $Q(\psi_j)$ from period s to t , $(\mu_{a|s} - \mu_{a|t})$, must be 0 since the population of employers is restricted to be identical in both periods. Similarly, the change in main effects of $Q(\tilde{\varepsilon}_{js})$, $(\mu_{c|s} - \mu_{c|t})$, must be 0 since the workforce productivity distribution is only measured at period s . The change in interaction of $Q(\psi_j)$ and $Q(\tilde{\varepsilon}_{js})$, $(\gamma_{ac|s} - \gamma_{ac|t})$, must also be 0 for the same reason.

This leaves two sets of parameters that are unconstrained by the null hypothesis—the change in main effects of $Q(\theta_i)$, $(\mu_{b|s} - \mu_{b|t})$, with $df = (\#Q(\theta_i) - 1)$ and the change in interaction of $Q(\psi_j)$ and $Q(\theta_i)$, $(\gamma_{ab|s} - \gamma_{ab|t})$, with $df = (\#Q(\psi_j) - 1) \times (\#Q(\theta_i) - 1)$. The parameters affected by the null hypothesis are the change in interaction of $Q(\theta_i)$ and $Q(\tilde{\varepsilon}_{js})$, $(\gamma_{bc|s} - \gamma_{bc|t})$, with $df = (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1)$ and the change in interaction of $Q(\psi_j)$, $Q(\theta_i)$ and $Q(\tilde{\varepsilon}_{js})$, $(\rho_{abc|s} - \rho_{abc|t})$, with $df = (\#Q(\psi_j) - 1) \times (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1)$. Under the null hypothesis $(\gamma_{bc|s} - \gamma_{bc|t}) = 0$ and $(\rho_{abc|s} - \rho_{abc|t}) = 0$ with $df = \nu_2 = (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1) + (\#Q(\psi_j) - 1) \times (\#Q(\theta_i) - 1) \times (\#Q(\tilde{\varepsilon}_{js}) - 1)$.

B.3.2 Computation of the Productive Workforce Test Statistics

We use the method of moments for test (B-3). The observations are firms j with positive employment in s . For each firm compute

$$x_j = \begin{bmatrix} \frac{n_{j1t}}{n_{j+t}} - \frac{n_{j1s}}{n_{j+s}} \\ \frac{n_{j2t}}{n_{j+t}} - \frac{n_{j2s}}{n_{j+s}} \\ \frac{n_{jt}}{n_{j+t}} - \frac{n_{js}}{n_{j+s}} \\ \dots \\ \frac{n_{j(\#Q(\theta_i)-1)t}}{n_{j+t}} - \frac{n_{j(\#Q(\theta_i)-1)s}}{n_{j+s}} \end{bmatrix}$$

where

$$n_{jqt} = \sum_{\{i|J(i,t)=j\}} 1(Q(\theta_i) = q).$$

and x_j is $[(\#Q(\theta_i) - 1) \times 1]$. For each value of a and c compute the vector of means and the covariance matrix

$$\bar{x}_{ac} = \frac{\sum_{\{j|Q(\psi_j)=a \wedge Q(\tilde{\epsilon}_{js})=c\}} n_{j+s} x_j}{\sum_{\{j|Q(\psi_j)=a \wedge Q(\tilde{\epsilon}_{js})=c\}} n_{j+s}}$$

$$V_{ac} = \frac{\sum_{\{j|Q(\psi_j)=a \wedge Q(\tilde{\epsilon}_{js})=c\}} n_{j+s} (x_j - \bar{x}_{ac})(x_j - \bar{x}_{ac})'}{\sum_{\{j|Q(\psi_j)=a \wedge Q(\tilde{\epsilon}_{js})=c\}} n_{j+s}}$$

$$N = \sum_j 1(j|\exists i : J(i, s) = j)$$

For each value of a compute the expected mean under the null hypothesis

$$\bar{x}_a = \frac{\sum_{\{j|Q(\psi_j)=a\}} n_{j+s} x_j}{\sum_{\{j|Q(\psi_j)=a\}} n_{j+s}}$$

Then,

$$X_{\nu_2}^2 = N \sum_{a,c} (\bar{x}_{ac} - \bar{x}_a)' V_{ac}^{-1} (\bar{x}_{ac} - \bar{x}_a).$$

Under the null hypothesis, $X_{\nu_2}^2$ follows a chi-square distribution with ν_2 degrees of freedom.

C Online Appendix: Posterior Distribution of the Parameter Vector

The posterior distribution of ρ given (Y, Z) is

$$p(\rho|Y, Z) \propto \mathcal{L}(\rho|Y, Z) \frac{1}{\sigma^{\nu_0+1}} \exp\left(-\frac{s_0^2}{\sigma^2}\right) \prod_{\ell=1}^L \pi_{a\ell}^{\frac{1}{L}-1} \prod_{m=1}^M \pi_{bm}^{\frac{1}{M}-1} \quad (\text{C-1})$$

$$\times \prod_{\ell=1}^L \prod_{m=0}^M \prod_{q=1}^Q \left(\pi_{q|\ell m}^{\frac{1}{Q}-1} \gamma_{\ell m q}^{\frac{1}{2}-1} (1 - \gamma_{\ell m q})^{\frac{1}{2}-1} \prod_{m'=0}^M \delta_{m'|\ell m q}^{\frac{1}{M+1}-1} \right).$$

This distribution factors into posterior distributions for the model parameters that are independent, conditional on the latent data, from which we sample.

To characterize these distributions, we introduce new notation. The matrix $G = [X A B K]$ is the full design of observed characteristics, ability, productivity, and match types given the observed and latent data. The term ν , which appears in the posterior of σ , is $\nu = N + \nu_0 - (L + M + Q)$. The sum of squared log earnings residuals is

$$s^2 = \frac{\left(\ln w - G \begin{bmatrix} \hat{\alpha} \\ \theta \\ \psi \\ \mu \end{bmatrix} \right)^T \left(\ln w - G \begin{bmatrix} \hat{\alpha} \\ \theta \\ \psi \\ \mu \end{bmatrix} \right)}{\nu}. \quad (\text{C-2})$$

The remaining parameters are sampled from Dirichlet posteriors, denoted by D.

Key to estimation are various counts from the completed data. $n_{a\ell}$ is the count of workers with ability type ℓ . n_{bm} is the number of employers in productivity type m . $n_{k|abq}$ is the number of matches observed in quality type q . n_{lmq}^{sep} is the number of observations in which a worker in ability type ℓ separates from an employer in productivity type m when match quality was q . Finally, $n_{m'\ellmq}^{trans}$ is the number of transitions by workers in ability type ℓ from a match with an employer in productivity type m and match quality type q to an employer in productivity type m' .

The posterior distribution of the wage equation parameters is

$$\begin{bmatrix} \alpha \\ \beta \\ \theta \\ \psi \\ \mu \end{bmatrix} | \sigma \sim N \left(\begin{bmatrix} \hat{\alpha} \\ \beta \\ \theta \\ \psi \\ \mu \end{bmatrix}, \sigma^2 (G^T G)^{-1} \right) \quad (\text{C-3})$$

where

$$\begin{bmatrix} \hat{\alpha} \\ \beta \\ \theta \\ \psi \\ \mu \end{bmatrix} = (G^T G)^{-1} G^T w,$$

and

$$\sigma^2 \sim \text{IG} \left(\frac{\nu}{2}, \frac{2}{\nu s^2} \right). \quad (\text{C-4})$$

The posterior distributions for the latent heterogeneity types are Dirichlet:

$$\pi_a \sim \text{D} \left(n_{a1} + \left(\frac{1}{L} \right), \dots, n_{aL} + \left(\frac{1}{L} \right) \right); \quad (\text{C-5})$$

$$\pi_b \sim \text{D} \left(n_{b1} + \left(\frac{1}{M} \right), \dots, n_{bM} + \left(\frac{1}{M} \right) \right); \quad (\text{C-6})$$

$$\pi_{k|ab} \sim \text{D} \left(n_{k|ab1} + \left(\frac{1}{Q} \right), \dots, n_{k|abQ} + \left(\frac{1}{Q} \right) \right). \quad (\text{C-7})$$

The posterior distributions of the separation and assignment parameters of the mobility model are also Dirichlet:

$$\gamma_{lmq} \sim \text{D} \left(n_{\ell mq}^{sep} + \left(\frac{1}{2} \right), n_{\ell mq}^{stay} + \left(\frac{1}{2} \right) \right); \quad (\text{C-8})$$

$$\delta_{b|lmq} \sim \text{D} \left(n_{0|\ell mq}^{trans} + \left(\frac{1}{M+1} \right), \dots, n_{M|\ell mq}^{trans} + \left(\frac{1}{M+1} \right) \right). \quad (\text{C-9})$$

D Online Appendix: The Mobility Model in Steady-State

The stationary distribution of the mobility model gives a steady-state distribution of employment spells across worker, employer, and match types. This, it turns out, is a model for the realized mobility network, characterized in the data by the design matrix of employer effects, F , and the associated cross-product term, $D^T F$. We also interpret it as a characterization of the selection model – the process by which particular matches are selected from the set of all possible matches.

The stationary distribution is simple to characterize: define $\lambda_{\ell,m,q}$ to be the expected number of matches in steady-state between workers of type ℓ and employers of type m on matches with quality q . Now define the diagonal matrix

$$\Lambda = \text{diag}([\lambda_{111}, \lambda_{112}, \dots, \lambda_{LMQ}]^T). \quad (\text{D-1})$$

Note that Λ does not account for transitions to non-employment. For exposition, suppose $L = M = Q = 2$ so Λ is 8×8 . In estimation, we let L , M , and Q vary and report results for the case $L = Q = M = 10$.

In steady-state, observed log earnings data $\ln w$ are drawn from a discrete distribution proportional to Λ . Net of the statistical residual, and the effect of observed time-varying characteristics, $X\beta$, the potential outcomes $\ln w - X\beta - \varepsilon$ are completely characterized by an $LMQ \times 1$ vector, \tilde{y} with

$$\tilde{y}_{\ell,m,q} = \alpha + \theta_{\ell} + \psi_m + \mu_q. \quad (\text{D-2})$$

The model therefore specifies

- Potential Outcomes: \tilde{y} , and
- Selection Process: Λ .

Define a set of indicator matrices analogous to the person, employer, and match design matri-

ces. For the $2 \times 2 \times 2$ model, this matrix is simply

$$[\tilde{D} \quad \tilde{F} \quad \tilde{G}] = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}. \quad (\text{D-3})$$

The notation \tilde{D}, \tilde{F} , and \tilde{G} highlights the connection between these reduced-dimension objects and the design matrices of worker and employer effects in the full data AKM model.

Net of $X\beta$ and ε , the earnings data are sampled from a distribution proportional to

$$\Lambda \tilde{y} = \Lambda \left(\tilde{D}\theta + \tilde{F}\psi + \tilde{G}\mu \right) \quad (\text{D-4})$$

and the full cross-product matrix is

$$[\tilde{D} \quad \tilde{F} \quad \tilde{G}]^T \Lambda [\tilde{D} \quad \tilde{F} \quad \tilde{G}] = \begin{bmatrix} \tilde{D}^T \Lambda \tilde{D} & \tilde{D}^T \Lambda \tilde{F} & \tilde{D}^T \Lambda \tilde{G} \\ \tilde{F}^T \Lambda \tilde{D} & \tilde{F}^T \Lambda \tilde{F} & \tilde{F}^T \Lambda \tilde{G} \\ \tilde{G}^T \Lambda \tilde{D} & \tilde{G}^T \Lambda \tilde{F} & \tilde{G}^T \Lambda \tilde{G} \end{bmatrix}. \quad (\text{D-5})$$

Notice that the upper-left block of the cross-product matrix in (D-5) is a model for the Laplacian of the realized mobility network, which is random noise around this steady-state distribution.

E Online Appendix: Estimation and Data Details

E.1 Parallelization of Employer Updates through Graph Coloring

To speed computation of the employer updates, we exploit the conditional independence restriction in the update formula, equation (13). For any employers, j and j' , we say j and j' are degree-one connected if any worker was observed to move from j directly to j' in the sample. The set, $N(j)$, is the set of all employers, j' , that are degree-one connected to j . Equation (13) implies that if j'' is not in $N(j)$ and j is not in $N(j'')$, then $\Pr[b_j = m | a, b_{-j}, k, Y, \rho]$ is independent of $\Pr[b_{j''} = m | a, b_{-j''}, k, Y, \rho]$ and, therefore, conditional on the rest of the latent data, the latent type of j and j'' can be updated at the same time (in parallel).

To fully exploit the network structure and conditional independence assumptions, we need groups of employers such that no two employers are degree-one connected. In the language of graph theory, this problem is equivalent to graph coloring in which the task is to color each node of a graph so that no two degree-one connected nodes have the same color, and to do so using the fewest colors possible.

For a general graph, the problem of finding the minimum number of colors is intractable. For

our task, it is sufficient to find a coloring that yields a small number of partitions relative to the highest degree node in the data (well over 1,000). To that end, we implement the greedy sequential coloring algorithm described in Gebremedhin et al. (2005). Briefly, the algorithm sorts network nodes from highest to lowest degree (that is, sorting employers in descending order by the number of job-to-job separations). The first node is assigned a color at random. For every other node, we assign the least frequent color that has not already been applied to one of its neighbors. If there is no such color, we add a new color to the list and continue.

In our data, this algorithm yields a coloring that partitions employers into 24 non-intersecting subsets. We update the employer types in parallel within each subset, and in sequence across the subsets. Our partition is well below the algorithmic worst-case guarantee: a coloring with as many colors as the highest-degree node in the graph, which is much greater than 1,000.

E.2 Calculation of Monte Carlo Standard Errors

When reporting results, we report Monte Carlo standard errors (MCSE) in place of, or in addition to, the posterior standard deviation. Unsurprisingly, we observe substantial autocorrelation across draws from the Gibbs sampler. The MCSE are computed using time-series methods that account for uncertainty about the location of the posterior distribution associated with autocorrelation in the chain. Using MCSE provides a practical and rigorous method for combining information across independent runs of the Gibbs sampler (we use three). The MCSE also fully exploit the information within each sample, while addressing within-thread autocorrelation, relative to more conventional ad hoc approaches like thinning the sample. Our ability to do so is all the more important given the computational burden of each draw. Even with the parallelization described in Section E.1, drawing from the Gibbs sampler is very time-consuming. Here, we describe implementation choices that affect our analysis. We refer the reader interested in the theoretical and practical details of computing the MCSE to the survey by Geyer (2011).

In calculating the MCSE, we implement the multivariate extension developed in Kosorok (2000) of initial sequence methods originally proposed by Geyer (1992). There are three variants of the initial sequence method, all of which exploit reversibility of the Markov Chain to determine the largest lag to include when computing the autocorrelation coefficient. The values reported in our tables are estimates from the initial positive sequence method, which are the most conservative. The other two methods, which we also implement, are the initial monotone and initial convex sequence methods. There is no meaningful difference across the estimates. In practice, we compute the univariate MCSE for each parameter due to numerical instability in the auto-covariance matrices.

E.3 Details of Variable Construction

Here we describe how the analysis variables are constructed from the LEHD microdata. Note that the raw UI records that supply the LEHD infrastructure are quarterly earnings records. Our analysis data are at the job-year level. Our key dependent variable, annual earnings, is constructed by summing all quarterly earnings records (converted to base year 2000 dollars using CPI-U) for

the same job (worker matched to firm) over the year in question. To deal with outliers in earnings, we Winsorize at the 0.01 and 99.99 percentiles.

Demographic characteristics of the worker are linked from the National Individual Characteristics File (NICF). These characteristics originate from Social Security records and other sources, including Census 2000 and the American Community Survey. These characteristics, and their construction, are described in Abowd et al. (2009), and subsequent internal research.

We also construct controls for the sequence of earnings records observed over the year. These controls address the problem that jobs that end mid-year will mechanically have lower earnings than jobs that last all year, even if the rate at which labor market earnings are acquired remains constant. These are based on a bit string, called `sixqwindow`, that records the observed pattern of quarters with positive earnings for a given job-year combination. The variable `sixqwindow` also records whether the worker was observed employed on the current job at the end of the preceding year and in the start of the subsequent year. So, for example, `sixqwindow=000011` for a job that started in the last quarter of the current year and continued into the first quarter of the next year. Likewise, `sixqwindow=110000` for a job on which a worker was employed at the end of the last year, and also reported earnings into the first quarter of the current year. As a final example, a job which is continuing from the preceding year, and continuing into the next year, and in which the worker was employed all year will have `sixqwindow=111111`.

There are 60 feasible values of `sixqwindow` (since strings of the form `x0000y` are ruled out). We summarize the salient information from these strings in a set of ten indicator variables. These are

- `sixq1` is an indicator equal to 1 if the sum of entries in `sixqwindow` is equal to 1, and zero otherwise.
- `sixq2` is an indicator equal to 1 if the sum of entries in `sixqwindow` is equal to 2, and zero otherwise.
- `sixq3--sixq6` are defined equivalently to the above.
- `sixqleft` is equal to 1 if `sixqwindow` has a continuous list of zeros from the right and ones from the left.
- `sixqright` is equal to 1 if `sixqwindow` has a continuous list of ones from the right and zeros from the left.
- `sixqinter` is equal to 1 if `sixqwindow` has a continuous list of ones interrupted by a single sequence of zeros.
- `sixq4th` is equal to 1 if `sixqwindow` indicates the worker was employed in the fourth quarter of the year.

These variables are effective in dealing with differences in job attachment over the year. Another option is to use the same information to convert the earnings data into an annualized level that measures the earnings a worker would earn if they accrued earnings at the same rate through the

entire year. The latter approach requires ad hoc assumptions about when jobs tend to end within the quarter. By contrast, we can see exactly how the endogenous mobility model treats such cases from the coefficient estimates in Table OA3. For example, consider the contrast between a job-year observation for which the worker is continuously employed relative to a job-year observations for which the worker is employed through the third quarter. In terms of the estimated parameters, the worker employed full time full year will have $sixq6=1$, $sixqlleft=1$, $sixqright=1$ (and all other job attachment dummies=0). By contrast, the worker employed through the second quarter will have $sixq3=1$ and $sixqlleft=1$ (and all other attachment dummies=0). Assuming all else is the same, the difference in predicted log earnings is 0.5321. That is, earnings on a job that ends during the third quarter are predicted to be 60 percent as large a job that lasts all year and continues into the next. The 60 percent figure captures the fact that jobs with reported earnings in the third quarter likely do not last to the end of the quarter.

E.4 Details of Selecting the Number of Latent Types

Following the guidance in Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2013, p. 536), our intention was to select a specification in which the number of latent types is an upper bound and let the data reveal the number of occupied types. To that end, through an initial model selection step, we selected the number of latent worker, firm, and match types to make the unexplained variation in the structural earnings model as close as possible to the residual variance from the AKM decomposition. We also favored specifications that are a priori symmetric in the number of latent types. This process resulted in our preferred specification with ten latent types for each dimension of heterogeneity. As discussed in the text, it appears ex post that while there is mass in each of the latent worker and match types, only four of the latent firm types have support. The missing categories are effectively collapsed in the posterior summaries.

An alternative procedure is to use a formal model selection criterion. For non-singular models like ours, the literature cautions that the Akaike, Bayesian, and Deviance Information Criteria are either technically infeasible, or not theoretically well-justified (Watanabe 2013; Gelman, Carlin, Stern, Dunson, Vehtari and Rubin 2013). The recommended alternative is the Watanabe-Akaike Information Criterion (WAIC) as it is both fully Bayesian, computationally tractable, and formally connected to cross-validation. See Gelman, Hwang and Vehtari (2013) for complete details.

We applied the WAIC to models with 3, 5, 7, and 10 latent types on each dimension. For the model selection exercise, we draw 1000 samples under each specification and compute the WAIC based on a 1-in-25 thinned subsample after a 500 sample burn-in. The results appear in Table OA6.

The WAIC is lowest for the model with 5 latent heterogeneity types and highest for our preferred model with 10 latent types. The model with 10 types has non-trivially higher likelihood than the others. These findings are consistent with our view that the model with 10 latent types represents an upper bound on the number for our purpose of fitting the existing data.

Figures OA3 and OA4 show the estimated wage components and distribution across latent types respectively for the models with 5, 7, and 10 latent types. These figures indicate that the pattern of model estimates is broadly consistent across specifications. It is only when we go to the model with 10 types that it becomes evident that several of the latent firm types are not filled. Again, these categories are effectively collapsed when the data are postprocessed to generate posterior

Table OA6: Likelihood and Watanabe-Akaike Information Criterion under Different Specifications

Model	WAIC	Likelihood
$L, M, Q = 3$	1, 174, 843	-431, 720
$L, M, Q = 5$	1, 109, 066	-363, 529
$L, M, Q = 7$	1, 139, 649	-367, 601
$L, M, Q = 10$	1, 153, 349	-337, 972

summaries. To instead perform an exhaustive model selection search using WAIC would require us to separately fit the model under each of the 1000 possible combinations of types. It is therefore infeasible. Given the goals of our analysis, the methodological literature and the data both support our choice to model 10 latent types as an explicit upper bound.

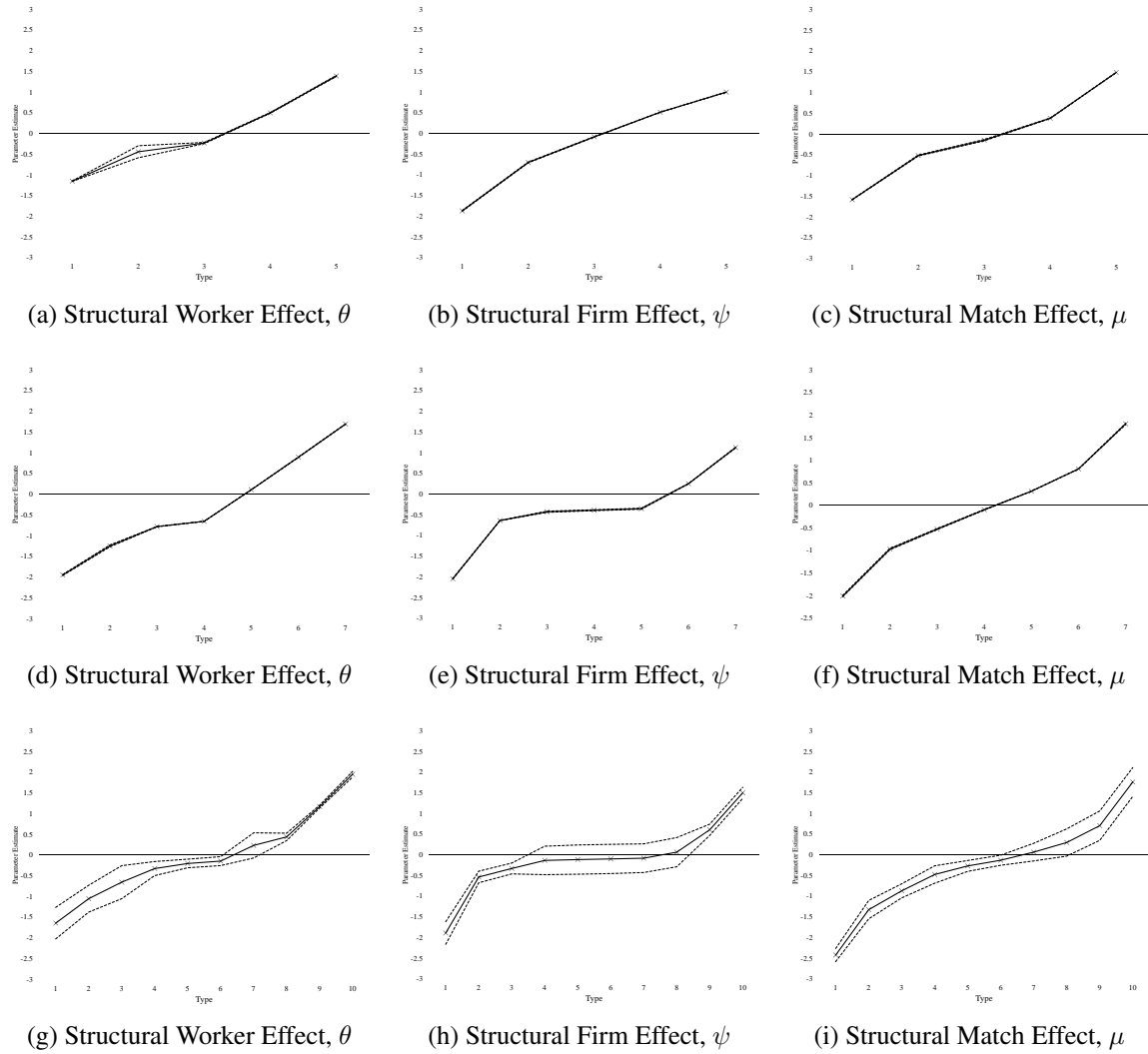
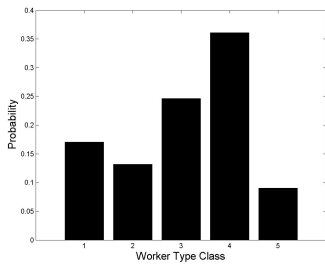
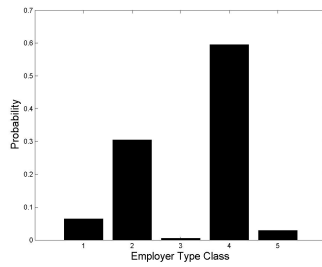


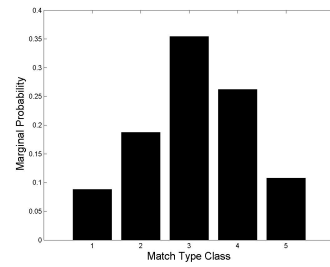
Figure OA3: Posterior distribution of earnings equation parameters. Dashed lines indicate $\pm 2 \times$ MCSE. The top row reports the specification with $L = M = Q = 5$ latent types. The second row reports the specification with $L = M = Q = 7$ latent types. The second row reports the specification with $L = M = Q = 10$ latent types.



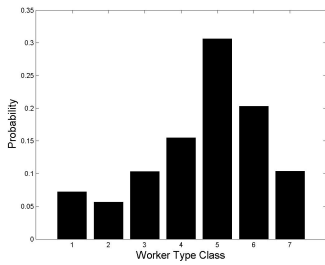
(a) Dist. of Worker Types



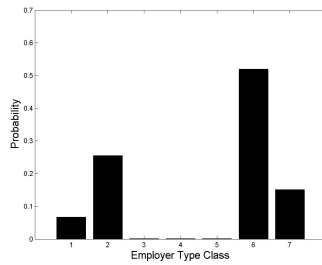
(b) Dist. of Employer Types



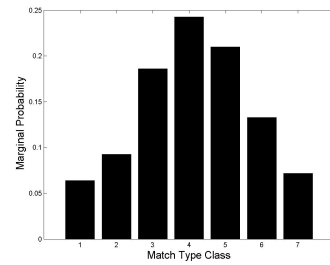
(c) Marginal Dist. of Match Types



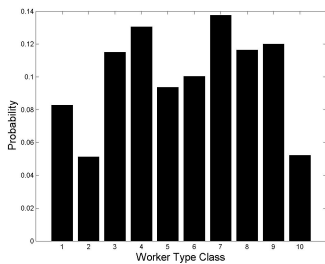
(d) Dist. of Worker Types



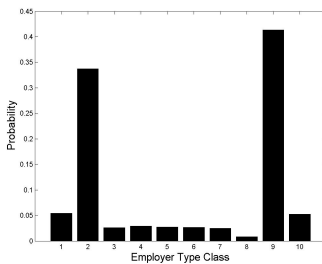
(e) Dist. of Employer Types



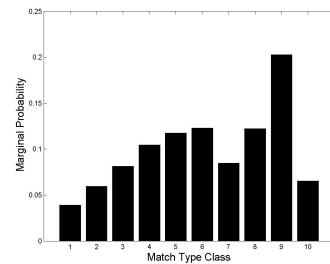
(f) Marginal Dist. of Match Types



(g) Dist. of Worker Types



(h) Dist. of Employer Types



(i) Marginal Dist. of Match Types

Figure OA4: Posterior distribution of workers, employers, and matches across latent types. The top row reports the specification with $L = M = Q = 5$ latent types. The second row reports the specification with $L = M = Q = 7$ latent types. The second row reports the specification with $L = M = Q = 10$ latent types.

F Online Appendix: Results using Extended Work Histories

We report results of estimating the structural model on an extended version of our main analysis sample. Specifically, we augment the main analysis sample by attaching the complete work history from 1990–2010 for each of the workers. This “extended work history” sample includes 1,778,490 person-year observations that cover 181,592 firms, and 389,718 matches. It is constructed to contain the primary analysis sample from the main text as a strict subset. The results reported here are based on 7,922 draws from three parallel runs of the Gibbs sampler after removing a 300 iteration burn-in. A complete archive of these results is available by request.

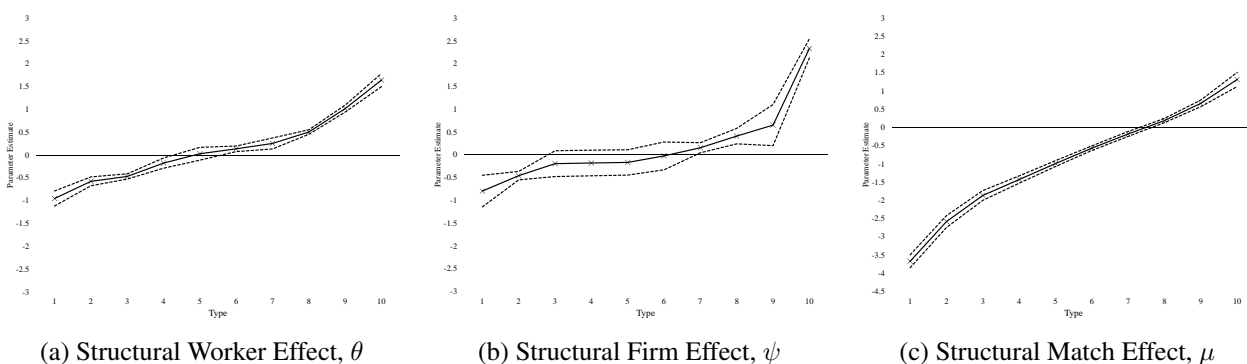


Figure OA5: Posterior distribution of earnings equation parameters. Dashed lines indicate the region within $2 \times MCSE$ of the estimate.

Table OA7 reports the posterior mean and MCSE of the parameter governing the population distribution of worker types, π_A , the population distribution of employer type, π_B , and the marginal probability for match type, π_K .

Table OA8 reports correlations, weighted by job duration, among earnings and its components as estimated by least squares (labeled AKM) and from our structural endogenous mobility model (labeled Gibbs). It is the analogue to Table 2 from the main text.

Table OA9 reports the results of estimating a firm-level regression of log revenue per worker onto the estimated firm effect, the average worker effect, and the average match effect. It is the analogue to Table 3 from the main text.

Table OA7: Posterior Distribution of Worker, Firm, and Match Population Heterogeneity: Extended Work Histories

	(1) Worker		(2) Firm			(3) Match	
	Mean	MCSE	Mean	MCSE		Mean	
π_{A1}	0.0875	0.0078	π_{B1}	0.2038	0.2668	π_{K1}	0.0093
π_{A2}	0.1317	0.0127	π_{B2}	0.5985	0.2682	π_{K2}	0.0249
π_{A3}	0.1448	0.0032	π_{B3}	0.0003	0.0001	π_{K3}	0.0916
π_{A4}	0.1379	0.0094	π_{B4}	0.0003	0.0001	π_{K4}	0.0616
π_{A5}	0.0988	0.0103	π_{B5}	0.0003	0.0001	π_{K5}	0.0509
π_{A6}	0.1005	0.0064	π_{B6}	0.0002	0.0001	π_{K6}	0.0620
π_{A7}	0.0835	0.0052	π_{B7}	0.0002	0.0000	π_{K7}	0.1014
π_{A8}	0.0695	0.0034	π_{B8}	0.0159	0.0114	π_{K8}	0.1572
π_{A9}	0.0968	0.0067	π_{B9}	0.0203	0.0129	π_{K9}	0.3463
π_{A10}	0.0490	0.0017	π_{B10}	0.1602	0.0049	π_{K10}	0.0948

Results from the structural model estimated using extended work histories. It is structured identically to Table 1 from the main text.

Table OA8: Correlation Matrix of Earnings Equation Parameters: Extended Work Histories

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	$\ln w$	$X\beta_{AKM}$	θ_{AKM}	ψ_{AKM}	μ_{AKM}	ε_{AKM}	$X\beta_{Gibbs}$	θ_{Gibbs}	ψ_{Gibbs}	μ_{Gibbs}	ε_{Gibbs}
$\ln w$	1.00										
$X\beta_{AKM}$	0.44	1.00									
θ_{AKM}	0.39	-0.49	1.00								
ψ_{AKM}	0.50	0.07	0.17	1.00							
μ_{AKM}	0.34	0.03	0.00	-0.00	1.00						
ε_{AKM}	0.20	-0.02	0.00	0.00	-0.00	1.00					
$X\beta_{Gibbs}$	0.78	0.56	0.25	0.24	0.04	-0.02	1.00				
θ_{Gibbs}	0.50	0.14	0.38	0.27	0.00	0.00	0.25	1.00			
ψ_{Gibbs}	0.27	0.02	0.12	0.42	0.11	0.00	0.10	0.04	1.00		
μ_{Gibbs}	0.06	0.05	-0.05	-0.10	0.28	0.00	-0.00	-0.23	-0.74	1.00	
ε_{Gibbs}	0.27	0.00	0.02	0.08	0.17	0.78	0.00	0.00	0.00	0.00	1.00

Table OA9: Regression of Log Revenue Per Worker on Structural and AKM Estimates of Earnings Decomposition Components: Extended Work Histories

	(1)		(2)	
	Structural		AKM	
	Coef	Ste	Coef	Ste
Firm Avg. θ	0.2288	0.0119	0.0234	0.0059
ψ	0.2431	0.0082	0.6735	0.0133
Firm Avg. μ	0.2046	0.009	0.0158	0.0094
Firm Avg. $X\beta$	0.0343	0.0063	-0.0231	0.0045
Intercept	3.4898	0.0532	3.9476	0.0247
N	60,116		60,116	