EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE   **Open Access**

A network theory of inter-firm labor flows

Eduardo López[1,2]* , Omar A. Guerrero[3,4] and Robert L. Axtell[1,3]

*Correspondence:
elopez22@gmu.edu
[1] Department of Computational and
Data Sciences, George Mason
University, 4400 University Drive, MS
6A2, Fairfax, VA, USA
[2] Green Templeton College,
University of Oxford, 43 Woodstock
Rd, OX2 6HG, Oxford, UK
Full list of author information is
available at the end of the article

Abstract

Using detailed administrative microdata for two countries, we build a modeling
framework that yields new explanations for the origin of firm sizes, the firm
contributions to unemployment, and the job-to-job mobility of workers between
firms. Firms are organized as nodes in networks where connections represent low
mobility barriers for workers. These labor flow networks are determined empirically,
and serve as the substrate in which workers transition between jobs. We show that
highly skewed firm size distributions are predicted from the connectivity of firms.
Further, our model permits the reconceptualization of unemployment as a local
network phenomenon related to both a notion of firm-specific unemployment and
the network vicinity of each firm. We find that firm-specific unemployment has a
highly skewed distribution. In coupling the study of job mobility and firm dynamics
the model provides a new analytical tool for industrial organization and makes it
possible to synthesize more targeted policies managing job mobility.

Keywords: Micro to macro models; Firm-size distribution; Employee mobility; Labor
flow networks

1 Introduction

The explanation of macro-phenomena on the basis of microscopic rules is a classic problem in social, economic and natural sciences [1]. Theoretical frameworks such as analytical sociology [2], economic microfoundations [3], and agent-based-modelling [4, 5] have emerged to address the need to connect the known rules of behavior at the individual level, along with interactions between those individuals, and the characteristics of the macro-behavior these rules generate. A distinctive element of micro to macro studies is their ability to address system heterogeneity [6, 7] where agent-based modelling (computational or otherwise) is especially useful due to its flexibility.

In the context of employment, the micro to macro problem is also present. However, its study has mostly been done through fundamentally macro-level approaches [8, 9] at the exclusion of much underlying micro-behavior. These approaches have utilized aggregation as one of their pillars, eliminating the numerous non-trivial and potentially dominant effects played by the ecology of heterogeneous employers, the firms [10, 11]. Most of the approaches that do incorporate some of the heterogeneity in the system [12, 13] do so by sectorizing the economy and coarsening the firm-level view (recent notable exceptions are [14–16] that incorporate firms).

Springer

Recently, through the use of microdata, the study of inter-firm networks of job-to-job mobility [14, 17, 18] (the phenomenon of an individual separating from one firm, potentially spending some unemployment time, and eventually reaching another firm) suggests that there is a rich set of phenomena taking place at the micro-level of the system that couples employers (firms) and their workers. Neither work on firm dynamics, nor work on employment had been able to identify or characterize such phenomena. Development of a new disaggregate model of the behavior of the firm-employee joint system, disciplined by microdata, is our main goal here.

Guided by micro-level empirical observations of the behavior of workers (which we also call agents) and the firms in which they work, we develop a model of job-to-job mobility that successfully establishes the micro-to-macro connection, empirically grounded at both levels. Our model is designed to understand the dynamics of agents in the firm landscape of large socioeconomic systems, and accurately predict large scale system behavior. We base our work on high resolution, extensive firm-employee matched records in two separate countries (Finland and Mexico), with data spanning large numbers of workers, firms, and multiple decades. The model we introduce is consistent with known key micro and macro-level regularities of the problem.

One of the main features of our model is that agents move between jobs inside an empirically constructed network of firms [14, 19]. This inter-firm network, acts as the substrate for workforce mobility. With this framework, we achieve some important results. First, we find that firm-size distributions are synthesized in the combination of the connectivity of firms in the firm network and the steady state movements of agents in the network. Furthermore, we propose a new notion of firm-specific unemployment, an intermediate state of agents between employment spells that offers a disaggregate picture of job-to-job mobility at the level of individual firms, and how each firm may contribute to unemployment. The use of a high resolution model to deal with job-to-job mobility is in itself an advance because it moves away from aggregate approaches to deal with the problem; aggregation destroys information about local effects like the specific dynamics of agent movements between close-neighbor firms in the network. Finally, our approach leads to the possibility to connect two areas of study that have traditionally been separate: job-to-job mobility and distribution of firm sizes.

We believe our framework has the great virtue of being applicable to a wide spectrum of situations, from calibrated models of true behavior, to idealized situations of academic or exploratory value. The reason for this flexibility is that the model captures and isolates the main ingredients of movement dynamics of workers and employers. These ingredients are: 1) the firm environment, where individuals travel as they change jobs, and 2) the rules of movement and behavior that individuals and firms satisfy. In this article, we choose those ingredients with enough realism to allow us to show proof of principle, i.e., to reproduce some important known facts about firms and worker dynamics at the best resolutions yet studied. However, the framework can be deployed with different environments and behavior to address many useful scenarios such as the potential consequences of policy interventions or the construction of baseline statistical models, to name a few.

It is important to emphasize that the dynamics that emerge from the coupled firm-employee system are of great practical interest. These dynamics have received attention in labour studies [8, 9], firm dynamics [20–23], and employment mobility [24, 25]. Usually, this research is parcelled into two separate problems: employment [8, 26] and

firms [23, 27, 28]. Such division reflects the absence of a theoretical framework capable of coupling them together. This division is mainly rooted on modelling choices. On one hand, employment models are often fully aggregate or very course-grained, eliminating the role of firm heterogeneity and the structure of job-to-job mobility. On the other, models of firm dynamics do not consider the reallocation of workers, discarding employment trajectories (e.g., Gibrat's law [29–31] focuses on the firm abstracted from its workers). Our model connects the labor mobility and firm dynamics literature, providing an integrated framework. Since our model is disaggregate and constructed to match data, it can serve as the starting point for more detailed work on dynamical aspects of employment and firms. Such work may be particularly useful to inform policy (specific measures with specific targets), provide better analysis of economic scenarios, and develop better forecasting frameworks.

2 Labor flow network

Evidence suggests that a worker's transition from a firm (i) to another firm (j) increases the probability of observing one or more subsequent worker transitions between the same firms i and j [18]. Such firm-to-firm transitions are the result of a confluence of factors (geography, social ties, skills, industry types, etc.) that affect choices made by individuals when navigating the employment landscape, and of firms when deciding to hire workers. Considerable research has been directed at the social ties factor [32–34] which has shown to be relevant but still constitutes just one out of numerous factors that play a large, if not larger roles.

A more data-driven and simultaneously disaggregate approach has been taken by Guerrero and Axtell [14] who proposed to represent firm-to-firm transitions with the use of a so-called *labor flow network* (LFN), where firms (nodes) are connected to one another if a job-to-job transition has been observed between them. Using administrative records from Finland and Mexico (see Sect. A.1), the authors constructed empirical networks, studied their structural properties, and found a highly heterogeneous firm environment with interesting regularities in the flow of workers such as virtually balanced flows in and out of individual firms. The advantage of the approach in Ref. [14] is that it takes into account all transitions regardless of their driving mechanism.

Our approach is to directly tackle the problem of modeling job-to-job transitions on the basis of empirically justified mechanisms, using as a starting point the evidence that previous job-to-job transitions may be predictive of new ones in the future, and that this justifies the use of inter-firm networks as part of the modeling framework for this problem. We model an agents-and-firms system by establishing: i) the LFN in which agents move on the basis of data, ii) a set of basic rules of how agents choose to navigate this network, and iii) how firms deal with the inflow and outflow of agents. In this section we focus on i), while ii) and iii) are developed in the next section.

Given available data for firms and agents, one would like to determine those *persistent* firm-to-firm transitions that, collected together as a set of edges interconnecting firms, constitute the substrate for a considerable fraction of the steady flow of the workforce. Intuitively, there is reason to believe in such a network (e.g., medical professionals move between hospitals, auto-technicians between car repair shops, the unemployed are more likely to search for a job where they live rather than far away, etc.), notions also captured

in work such as [13, 35]. Note that persistent transitions justify the use of labor flow networks. If transitions were a consequence of simple random events with no repeated transitions other than by chance, then there would be no need for a network: a transition from one node would simply be a random jump to any other node (this is a view consistently presented in current literature on aggregate studies in labor economics [8]). Transitions for which one cannot build evidence that they will occur again in the future are labelled random.
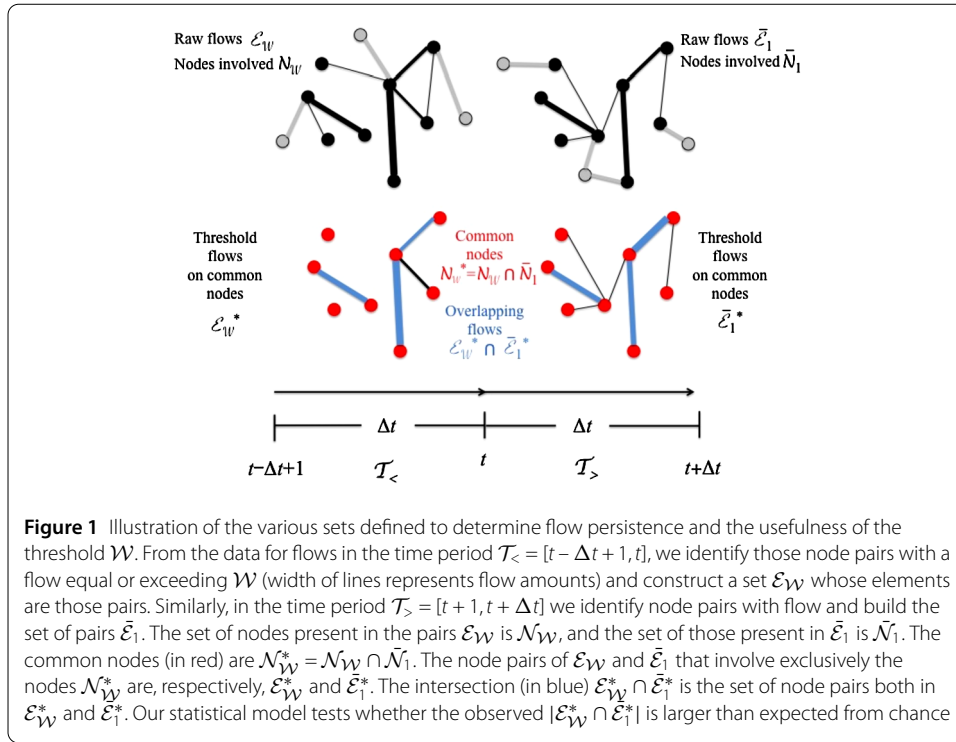
To understand persistent transitions and their consequences better, we proceed as follows.

2.1 Testing flow persistence between firms

The work by Collet and Hedström [18] focused on virtually comprehensive data for Stockholm found that the probability that two firms i and j experience new worker transitions between them after having had at least one prior transition is approximately 1027 times larger than transitions between any two random firms in the same system. The implication of this is that observing one transition between firms is a strong indicator of future transitions and thus, a strong suggestion that worker movements can be reliably modelled by introducing links between firms that exhibit certain flow amounts between them. In this article, we perform three[a] related tests (the last one is mainly contained in Sect. A.3.3) to confirm that employee transitions between firms have the same temporal predictability reported in Ref. [18]. Two of our tests are designed to adjust for firm sizes (an improvement over the methods in [18]); our results support the notion that past transitions are predictive of future ones.

We focus on the Finnish data. Consider worker flows in a window of time of $2\Delta t$ years centered around a given year t. Loosely speaking, we want to test for every pair of firms with flows in the first Δt years, from $t - \Delta t + 1$ to t, if they also tend to have flows in the second Δt years from $t + 1$ to $t + \Delta t$, above what would be expected from chance. We label the two time periods of size Δt before and after t as $\mathcal{T}_< = [t - \Delta t + 1, t]$ and $\mathcal{T}_> = [t + 1, t + \Delta t]$. To be specific, we define $f_{ij}(t)$ as the number of individuals changing jobs from firm i to firm j in year t, $F_{ij}(t) = f_{ij}(t) + f_{ji}(t)$ the total undirected flow between i and j in t, $\mathcal{F}_{ij}(t, \Delta t) = \sum_{t'=t-\Delta t+1}^{t} F_{ij}(t')$ the total flow between a pair ij during $\mathcal{T}_<$, and $\bar{\mathcal{F}}_{ij}(t, \Delta t) = \sum_{t'=t+1}^{t+\Delta t} F_{ij}(t')$ the flow between ij during $\mathcal{T}_>$. Let us introduce an arbitrary threshold $\mathcal{W}$ for $\mathcal{F}_{ij}$ such that, if the pair of nodes i and j satisfy it ($\mathcal{F}_{ij} \geq \mathcal{W}$), we check if the pair ij has flow $\bar{\mathcal{F}}_{ij} > 0$. In other words, we require a minimum flow amount $\mathcal{W}$ to track a pair of nodes from $\mathcal{T}_<$ during $\mathcal{T}_>$. We then test how informative the value of $\mathcal{W}$ is when trying to forecast which node pairs have flow in $\mathcal{T}_>$. If $\mathcal{W}$ successfully helps forecast active node pairs in the future, then we can treat $\mathcal{W}$ as an *acceptance criterion* for adding a link between i and j, thus generating a *labor flow network.*

To determine if the criterion is useful, we define several quantities. First, we capture the set of flows $\mathcal{E}(t, \Delta t, \mathcal{W}) = \{(ij)|\mathcal{F}_{ij}(t, \Delta t) \geq \mathcal{W}\}$, i.e. the pairs of nodes during $\mathcal{T}_<$ that achieve threshold $\mathcal{W}$. Together with these node pairs, we collect the unique nodes $\mathcal{N}(t, \Delta t, \mathcal{W}) = \{q|(q = i \vee q = j) \wedge (i, j) \in \mathcal{E}(t, \Delta t, \mathcal{W})\}$, that is, all the nodes that take part in any of the node pairs in $\mathcal{E}(t, \Delta t, \mathcal{W})$. We also define a set $\bar{\mathcal{E}}(t, \Delta t, 1) = \{(ij)|\bar{\mathcal{F}}_{ij}(t, \Delta t) \geq 1\}$, the node pairs that have at least one flow event during $\mathcal{T}_>$. Associated with this, we define $\bar{\mathcal{N}}(t, \Delta t, \mathcal{W} = 1) = \{q|(q = i \vee q = j) \wedge (i, j) \in \bar{\mathcal{E}}(t, \Delta t, 1)\}$, the nodes that take part in $\bar{\mathcal{E}}(t, \Delta t, 1)$. Since the nodes in $\mathcal{N}(t, \Delta t, \mathcal{W})$ or in $\bar{\mathcal{N}}(t, \Delta t, 1)$ may not all be present through

Figure 1 Illustration of the various sets defined to determine flow persistence and the usefulness of the threshold $\mathcal{W}$. From the data for flows in the time period $\mathcal{T}_< = [t - \Delta t + 1, t]$, we identify those node pairs with a flow equal or exceeding $\mathcal{W}$ (width of lines represents flow amounts) and construct a set $\mathcal{E}_{\mathcal{W}}$ whose elements are those pairs. Similarly, in the time period $\mathcal{T}_> = [t + 1, t + \Delta t]$ we identify node pairs with flow and build the set of pairs $\bar{\mathcal{E}}_1$. The set of nodes present in the pairs $\mathcal{E}_{\mathcal{W}}$ is $\mathcal{N}_{\mathcal{W}}$, and the set of those present in $\bar{\mathcal{E}}_1$ is $\bar{\mathcal{N}}_1$. The common nodes (in red) are $\mathcal{N}_{\mathcal{W}}^* = \mathcal{N}_{\mathcal{W}} \cap \bar{\mathcal{N}}_1$. The node pairs of $\mathcal{E}_{\mathcal{W}}$ and $\bar{\mathcal{E}}_1$ that involve exclusively the nodes $\mathcal{N}_{\mathcal{W}}^*$ are, respectively, $\mathcal{E}_{\mathcal{W}}^*$ and $\bar{\mathcal{E}}_1^*$. The intersection (in blue) $\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*$ is the set of node pairs both in $\mathcal{E}_{\mathcal{W}}^*$ and $\bar{\mathcal{E}}_1^*$. Our statistical model tests whether the observed $|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|$ is larger than expected from chance
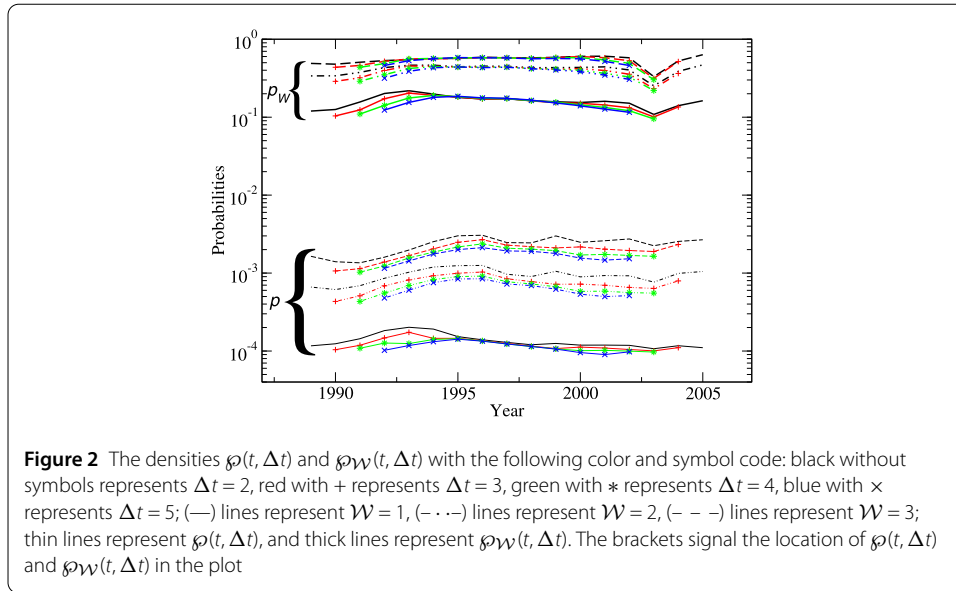
the entirety of the time period between $t - \Delta t + 1$ and $t + \Delta t$, we also define the intersection $\mathcal{N}^*(t, \Delta t, \mathcal{W}) = \mathcal{N}(t, \Delta t, \mathcal{W}) \cap \bar{\mathcal{N}}(t, \Delta t, 1)$, composed of nodes that are present in both $\mathcal{E}(t, \Delta t, \mathcal{W})$ and $\bar{\mathcal{E}}(t, \Delta t, 1)$, and also the sets $\mathcal{E}^*(t, \Delta t, \mathcal{W})$ and $\bar{\mathcal{E}}^*(t, \Delta t, 1)$ which are, respectively, the subsets of $\mathcal{E}(t, \Delta t, \mathcal{W})$ and $\bar{\mathcal{E}}(t, \Delta t, 1)$ that only include node pairs where both nodes are in $\mathcal{N}^*(t, \Delta t, \mathcal{W})$. This guarantees that comparisons between flows before and after year t are well defined. To avoid cumbersome notation, from here we use the shortened notation $\mathcal{E}_{\mathcal{W}}^* \equiv \mathcal{E}^*(t, \Delta t, \mathcal{W})$, $\bar{\mathcal{E}}_1^* \equiv \bar{\mathcal{E}}^*(t, \Delta t, 1)$, and $\mathcal{N}_{\mathcal{W}}^* \equiv \mathcal{N}^*(t, \Delta t, \mathcal{W})$ (an illustrative diagram of the process can be found in Fig. 1).

If the information of which pairs "are connected" in $\mathcal{E}_{\mathcal{W}}^*$ allows us to predict to some extend the flows in $\bar{\mathcal{E}}_1^*$, then it means that the criterion used to create $\mathcal{E}_{\mathcal{W}}^*$ is informative. To test if there is a relationship between $\mathcal{E}_{\mathcal{W}}^*$ and $\bar{\mathcal{E}}_1^*$, we determine the number of firm-pair flows in the period $\mathcal{T}_>$ that were also firm-pair flows ($\geq \mathcal{W}$) in the years $\mathcal{T}_<$; to obtain a fraction, we divide the number by $|\mathcal{E}_{\mathcal{W}}^*|$. This produces the density $\wp_{\mathcal{W}}(t)$, defined as

$$\wp_{\mathcal{W}}(t, \Delta t) = \frac{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|}{|\mathcal{E}_{\mathcal{W}}^*|}. \tag{1}$$

The denominator counts the opportunities for a node pair that satisfies the flow threshold during $\mathcal{T}_<$ to also have flow subsequently; the numerator counts how many times the opportunities are actually realized. Note that if the fraction took the value of 1 it would mean that all pairs that had flows of magnitude $\geq \mathcal{W}$ up to year t also had flows after year t; if the value were 0, it would mean that none of the pairs with flows $\geq \mathcal{W}$ up to year t had flows after that year. Crucially (establishing a null model, developed in detail in Sect. A.3.1), if the flows captured in $\mathcal{E}_{\mathcal{W}}^*$ provided no relevant information about the flows captured in

Figure 2 The densities $\wp(t, \Delta t)$ and $\wp_{\mathcal{W}}(t, \Delta t)$ with the following color and symbol code: black without symbols represents $\Delta t = 2$, red with + represents $\Delta t = 3$, green with ∗ represents $\Delta t = 4$, blue with × represents $\Delta t = 5$; (——) lines represent $\mathcal{W} = 1$, (– · ·–) lines represent $\mathcal{W} = 2$, (– – –) lines represent $\mathcal{W} = 3$; thin lines represent $\wp(t, \Delta t)$, and thick lines represent $\wp_{\mathcal{W}}(t, \Delta t)$. The brackets signal the location of $\wp(t, \Delta t)$ and $\wp_{\mathcal{W}}(t, \Delta t)$ in the plot

$\bar{\mathcal{E}}^*(t, \Delta t, 1)$, one would expect $\wp_{\mathcal{W}}(t, \Delta t)$ to be the same as
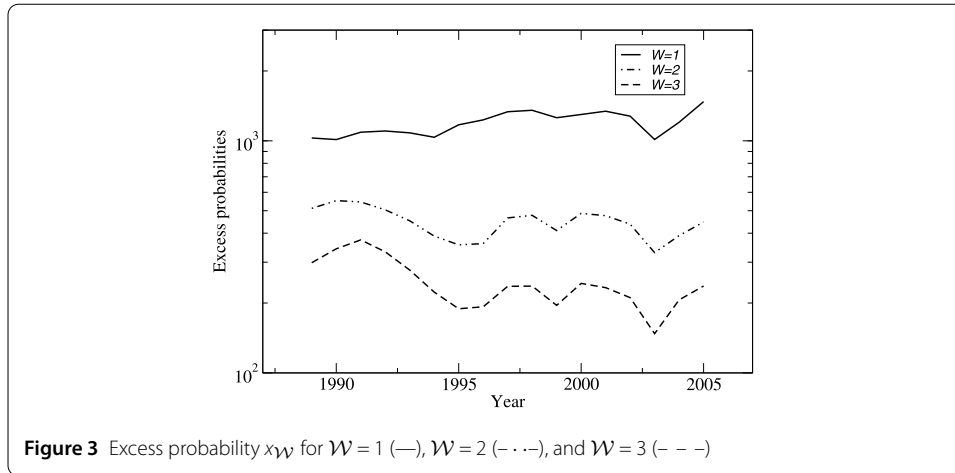
$$\wp(t, \Delta t) = \frac{|\bar{\mathcal{E}}_1^*|}{\binom{|\mathcal{N}_{\mathcal{W}}^*|}{2}}, \tag{2}$$

the density of node pairs with a flow of at least 1 during $\mathcal{T}_>$. The numerator in Eq. (2) counts the number of node pairs with flow and the denominator the total number of possible node pairs (this equation is the usual link density equation of a network). An intuitive way to understand why $\wp$ and $\wp_{\mathcal{W}}$ should be the same when the flows before and after t are not related is to note that by pure chance the expectation $\langle |\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| \rangle$ of the number of times a node pair in $\mathcal{E}_{\mathcal{W}}^*$ is also in $\bar{\mathcal{E}}_1^*$ is given by $\wp|\mathcal{E}_{\mathcal{W}}^*|$, i.e. the number of trials $|\mathcal{E}_{\mathcal{W}}^*|$ of finding a connection times the success rate $\wp$; if we insert this expectation into the numerator of Eq. (1), we confirm our claim. In fact, the null model is described by a hypergeometric distribution for which $\wp$ is the expected value for $\wp_{\mathcal{W}}$ (in Sect. A.3.1 we explain this and estimate the p-value which confirms the significance of our threshold test). If the flows $\mathcal{E}_{\mathcal{W}}^*$ are indicative of the flows $\bar{\mathcal{E}}_1^*$, we expect $\wp_{\mathcal{W}} > \wp$. To test this, we calculate the excess probability $x_{\mathcal{W}}$, given by the ratio

$$x_{\mathcal{W}}(t, \Delta t) = \frac{\wp_{\mathcal{W}}(t, \Delta t)}{\wp(t, \Delta t)} \tag{3}$$

that highlights any potential increase of $\wp_{\mathcal{W}}$ over $\wp$.

We compute $\wp(t, \Delta t), \wp_{\mathcal{W}}(t, \Delta t)$, and $x_{\mathcal{W}}(t, \Delta t)$ for a combination of years t and time windows Δt to get a comprehensive picture of the situation (see Sect. A.1 for details on data). Note that the size of the time window Δt impacts the range of years used: since the earliest dataset is for 1988 and the latest for 2007, then a window of time Δt means the years of analysis are $1987 + \Delta t \leq t \leq 2007 - \Delta t$. In Fig. 2 we present results for thresholds $\mathcal{W} = 1, 2, 3$, and for time windows $\Delta t = 2, 3, 4, 5$ (additional analysis and robustness checks are shown in Sect. A.7). Clearly, $\wp_{\mathcal{W}} \gg \wp$ by orders of magnitude across all combinations
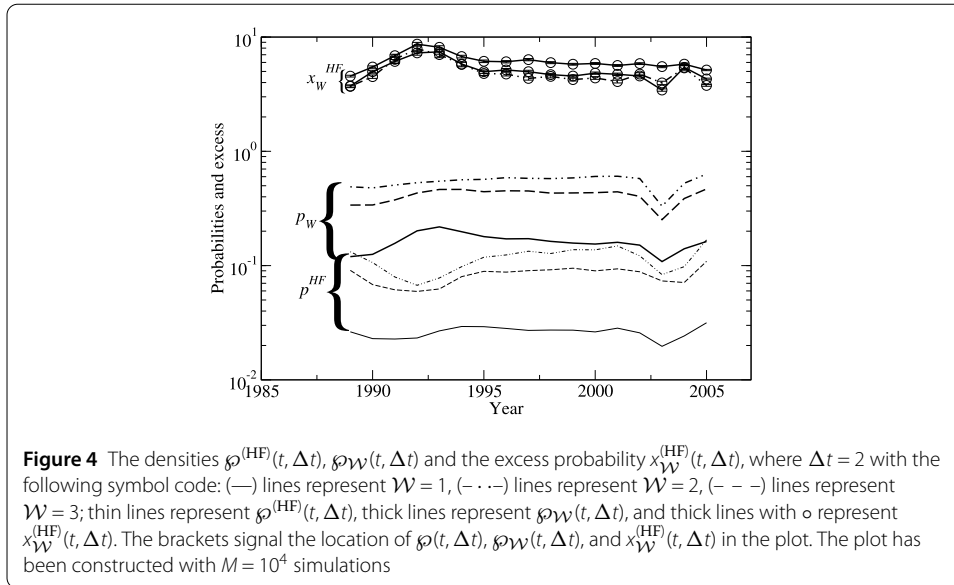
Figure 3 Excess probability $x_{\mathcal{W}}$ for $\mathcal{W} = 1$ (—), $\mathcal{W} = 2$ ($- \cdot \cdot -$), and $\mathcal{W} = 3$ ($- - -$)

of $\mathcal{W}, \Delta t$. This is because the observed values of $|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|$ are indeed also orders of magnitude larger than their expectations $\langle |\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| \rangle = \wp |\mathcal{E}_{\mathcal{W}}^*|$ (see Table 3). The size of the time window Δt does not have a pronounced effect on any of the probabilities, although $\wp$ seems slightly more sensitive to it than $\wp_{\mathcal{W}}$. On the other hand, $\mathcal{W}$ increases all the probabilities in the plot. More importantly, specifically for $\wp_{\mathcal{W}}$ we can see that as $\mathcal{W}$ increases from 1 to 3, the likelihoods of any of the flows in $\mathcal{E}_{\mathcal{W}}^*$ to also appear in $\bar{\mathcal{E}}_1^*$ increase from over ($\wp_{\mathcal{W}} \approx$) 10% to around ($\wp_{\mathcal{W}} \approx$) 50–60% and in some years even more.[b]

The excess probabilities for some of the previous curves are presented next, restricted only to $\Delta t = 2$ given the minor impact of Δt on the results, but including the thresholds $\mathcal{W} = 1, 2, 3$ (see Fig. 3). For $\mathcal{W} = 1$, the excess probability is largest, which means that the first flow event generates the greatest fractional increase between $\wp$ and $\wp_{\mathcal{W}}$ (an increase of more than 10^3, the same order of magnitude as in [18]). The threshold $\mathcal{W} = 2$ has an excess probability in the 400 to 500 range. Threshold $\mathcal{W} = 3$ has excess probabilities in the range of 150 to 350. The decrease of the, still considerably large, excess probabilities $x_{\mathcal{W}}$ with $\mathcal{W}$ is a consequence of the saturation effect that $\wp_{\mathcal{W}}$ undergoes (as it approaches 1 with increasing $\mathcal{W}$) in comparison to the unsaturated evolution of $\wp$ with $\mathcal{W}$.

To improve on the previous test, we present a modification that addresses the heterogeneity of firm sizes. Concretely, note that if flow occurs between two large firms within the time interval $\mathcal{T}_<$, then just by random chance the likelihood that another flow event occurs between these firms in the interval $\mathcal{T}_>$ should be larger than when the same situation occurs between two small firms. Thus, one could wonder if $\wp_{\mathcal{W}} \gg \wp$ might break down due to the firm-size heterogeneity.

The modification we introduce concerns a change to the null model. In the previous test, the null model predicts that $\wp_{\mathcal{W}}$ should approach the value of $\wp$, but is instead found to be orders of magnitude larger. That test rests on the assumption that any of the flows in $\bar{\mathcal{E}}_1^*$ are equally likely to occur among any pair of nodes in $\mathcal{N}_{\mathcal{W}}^*$, despite the differences in those nodes.

The new null model is constructed via Monte Carlo by randomizing the flows contained in $\bar{\mathcal{E}}_1^*$ among the nodes $i \in \mathcal{N}_{\mathcal{W}}^*$ with each node i preserving its observed in- and out-flows. For each random rewiring of $\bar{\mathcal{E}}_1^*$ we obtain a set of simulated flows $\bar{\mathcal{S}}_1^*$, and determine the size $|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|$ of their overlap with the flows for the period $\mathcal{T}_<$. For M realizations, we can calculate an expectation value $\langle |\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*| \rangle_{\text{(HF)}}$ where HF stands for heterogeneous firms.

Figure 4 The densities $\wp^{(\mathrm{HF})}(t, \Delta t)$, $\wp_{\mathcal{W}}(t, \Delta t)$ and the excess probability $x_{\mathcal{W}}^{(\mathrm{HF})}(t, \Delta t)$, where $\Delta t = 2$ with the following symbol code: (—) lines represent $\mathcal{W} = 1$, (– · ·–) lines represent $\mathcal{W} = 2$, (– – –) lines represent $\mathcal{W} = 3$; thin lines represent $\wp^{(\mathrm{HF})}(t, \Delta t)$, thick lines represent $\wp_{\mathcal{W}}(t, \Delta t)$, and thick lines with ○ represent $x_{\mathcal{W}}^{(\mathrm{HF})}(t, \Delta t)$. The brackets signal the location of $\wp(t, \Delta t)$, $\wp_{\mathcal{W}}(t, \Delta t)$, and $x_{\mathcal{W}}^{(\mathrm{HF})}(t, \Delta t)$ in the plot. The plot has been constructed with $M = 10^4$ simulations

In a similar way to the previous test, we compare $\wp_{\mathcal{W}}$ against the average density

$$\wp^{(\mathrm{HF})} = \left\langle \wp(t, \Delta t) \right\rangle_{\mathrm{HF}} = \frac{\langle |\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*| \rangle_{(\mathrm{HF})}}{|\mathcal{E}_{\mathcal{W}}^*|}, \tag{4}$$

as well as the excess probability $x^{(\mathrm{HF})} = \wp_{\mathcal{W}}/\wp^{(\mathrm{HF})}$. The results are shown in Fig. 4. In a similar way to the previous test, we find that $\wp_{\mathcal{W}} \gg \wp^{(\mathrm{HF})}$. The excess probability $x_{\mathcal{W}}^{(\mathrm{HF})}$ has values that roughly correspond to a factor of 10, not as large as $x_{\mathcal{W}}$ but still significant; the fact that $x_{\mathcal{W}}^{(\mathrm{HF})} < x_{\mathcal{W}}$ provides an *a posteriori* justification for the need for this new test. Additional considerations, including a discussion of *p*-values can be found in Sect. A.3.2.

As a final check, we present in Sect. A.3.3 an additional test that focuses on the amount of flow predicted by our threshold method in comparison with the heterogeneous firm null model. We find that the node pairs identified by the threshold method carry anywhere from about 30% to as much as 70% of the overall flow in the network depending on the specific t. The flows carried by the node pairs emerging from the null model consistently carry about a tenth of this flow.

Summarizing the results of the previous three tests, it is clear that the use of a threshold $\mathcal{W}$ identifies flows that persist into the future, and carry a very large fraction of the overall flow of the system. Furthermore, the likelihood that these flows are seen again increases monotonically with $\mathcal{W}$. For $\mathcal{W} = 2$, we already find that the likelihood of a pair of nodes having repeated flows is around ($\wp_{\mathcal{W}=2} \approx$) 40%. As we argue in Sect. A.7, the qualitative results of our analyses do not change by increasing this parameter, and therefore, in the main article we present results with $\mathcal{W} = 2$ which balances size of the sample with a significant certainty that flows are persistent.

2.2 Assembling the labor flow network

We construct the LFN G for a given dataset (Finland or Mexico) by assembling all N firms together with the edges that are found to be persistent according to the criterion above. The network is unweighted and undirected, characterized by the symmetric adjacency matrix $\mathbf{A}$ of dimension $N \times N$, with $\mathbf{A}_{ij} = \mathbf{A}_{ji} = 1$ if i and j are connected and zero otherwise.

In order to make better use of the data, we construct a network on the basis of the entire time frame for each dataset (20 years for Finland, and just over 29 years for Mexico). This is supported by results in Sect. A.8, specially those associated with Fig. 24. The LFNs built by this procedure ($\mathcal{W} = 2$) are found to carry a large portion of the job-to-job transitions: in Finland $\approx 60.33\%$ out of the total of 1,808,412 transitions observed, and in Mexico, $\approx 33.7\%$ out of the total of 624,880 transitions. Extending the criterion to include edges where transitions occur 3 times or more ($\mathcal{W} = 3$), the number of transitions captured is still high, with $\approx 51.17\%$ for Finland, and $\approx 24.8\%$ for Mexico.

The network that results from this procedure is characterized by a skew distributions of degree k_i (number of neighbors of i), total transitions through each node τ_i (typically called node strength in the networks literature), and link weights F_{ij}. Beyond the non-trivial nature of the distributions, it is worth mentioning that each of the quantities is necessary for a full description of the network: for instance, it is not enough to know the strength of a node τ (or even its directed versions $\tau^{(in)}, \tau^{(out)}$) in order to know k. This is important to realize as it indicates that the network structure or the results we describe later *do not* emerge directly from a single mechanism such as labor supply (which would imply that τ statistically describes k). The distributions, along with a lengthy discussion of their interpretation, are shown in the Appendix, Sect. A.4.

As a final point in this section, let us discuss the unweighted undirected nature of G. Qualitatively, LFNs are defined to reflect the presence (or absence when there is no edge) of employment "affinity" between two firms, necessary for firm-to-firm transitions. This approach captures the notion of a categorical relationship between firms [36]. Considering the limited microscopic data available, this is the most unbiased choice we can make when modelling observed persistent transitions between firms. If the choice is sound, the model should be able to accurately reproduce observations, as we confirm in this article. The choice of considering an edge as a categorical relationship also leads to the undirected assumption, as there is no *a priori* reason to discard transitions in either direction, and since the weights are all equal, there is no need to have directed edges.

3 Modelling firms and individuals

Having established that LFNs capture a large number of job-to-job transitions, we proceed to model agents navigating these networks.

Given that the specific choices made by agents travelling in the LFN are likely to be partially driven by chance, we introduce a discrete time stochastic model that encodes the behavior of agents and firms in the network.[c] The rules of the model are given by the following:

1. *Agents*: These can have two states, employed or unemployed. At the start of the time step, an employed agent remains in its firm, say i, with probability $1 - \lambda_i$, and leaves (or separates) with probability λ_i. An unemployed agent at the start of the time step first determines whether neighbors of firm i are accepting applications and, if any of them are, chooses one of them to apply to with uniform probability.

2. *Firms*: At the start of the time step, firms make a choice to receive applications with probability v_i or not receive them with probability $1 - v_i$. If firms choose to receive applicants, then each is accepted with a probability h_i.

The set of neighbors of i is denoted by Γ_i, and has size $k_i = |\Gamma_i|$, the degree of i. A specific subset of neighbors of node i receiving applicants in a given time step is denoted by γ_i,

and this can change from step to step. In this time step, the number of open neighbors is $|\gamma_i|$. Our calculation contains terms for configurations for which we condition γ_i to have neighbor j open, and in those cases the configuration is denoted by $\gamma_i^{(j)}$. The occurrence of configuration γ_i is a random event with probability $\Pr(\gamma_i) = \prod_{j \in \gamma_i} v_j \prod_{\ell \in \Gamma_i \setminus \{\gamma_i\}} (1 - v_\ell)$. We also find it useful to define the average hiring rate $\langle h \rangle_{\gamma_i} = \sum_{j \in \gamma_i} h_j / |\gamma_i|$ of neighbors of i given γ_i.

To track the state of the system, we define the probabilities $r(i, t)$ and $s(i, t)$ that an agent is, respectively, employed or unemployed at node i at time t. These two probabilities, explained in detail below, satisfy the equations

$$r(i,t) = (1 - \lambda_i) r(i, t-1) + h_i \sum_{j \in \Gamma_i} s(j, t-1) \sum_{\{\gamma_j^{(i)}\}} \frac{1}{|\gamma_j^{(i)}|} \Pr(\gamma_j^{(i)}) \tag{5}$$

and

$$s(i,t) = \lambda_i r(i, t-1)$$
$$+ s(i, t-1) \left[\sum_{\{\gamma_i \neq \emptyset\}} \Pr(\gamma_i) \frac{1}{|\gamma_i|} \sum_{j \in \gamma_i} (1 - h_j) + \Pr(\gamma_i = \emptyset) \right]. \tag{6}$$

The first equation states that the probability for an agent to be employed at node i at time t is given by the probability to be employed at node i at time $t-1$ and to not separate, plus the probability that the agent is unemployed at one of the neighbors of i, that i is accepting applications, that the agent chooses to apply to i, and that the application by the agent leads to being hired. The second equation states that the probability to be unemployed at i at time t is given by the probability to be employed at i at time $t-1$ and be separated, or to have been unemployed at time $t-1$ at i but not find a job among the neighbors of i, either because none of them are receiving applications, or because the agent chooses to apply to one of the neighbors and is not hired. Note that the brackets in Eq. (6) simplify to $1 - \sum_{\{\gamma_i \neq \emptyset\}} \langle h \rangle_{\gamma_i} \Pr(\gamma_i)$.

The rules and equations above (Eqs. (5) and (6)) give rise to a Markov chain, where workers act independently of one another, and the model parameters are static in time. We assume that the parameters h_i and v_i are independent of k_i, and that $0 < h_i, v_i \leq 1$. With these assumptions, and by calibrating from data the remaining model parameters, namely k_i and λ_i, we show below that the model reproduces key empirical observations. To explore the model, we now present an analysis of its steady state for the cases of heterogeneous v_i (each firm has its on value of v) and homogeneous $v_i = v$ (constant v).

In the steady state, the conditions $r(i, t) - r(i, t-1) = 0$ and $s(i, t) - s(i, t-1) = 0$ are satisfied. Using the notation $r(i, t) \to r_\infty(i)$ and $s(i, t) \to s_\infty(i)$ we find

$$0 = -\lambda_i r_\infty(i) + h_i \sum_{j \in \Gamma_i} s_\infty(j) \sum_{\{\gamma_j^{(i)}\}} \frac{1}{|\gamma_j^{(i)}|} \Pr(\gamma_j^{(i)}), \tag{7}$$

$$0 = \lambda_i r_\infty(i) - s_\infty(i) \sum_{\{\gamma_i \neq \emptyset\}} \langle h \rangle_{\gamma_i} \Pr(\gamma_i), \tag{8}$$

which, when solved for $r_\infty(i)$, lead to the matrix equation

$$\mathbf{\Lambda X} = \mathbf{0}, \tag{9}$$

where $\mathbf{\Lambda}$ is an $N \times N$ matrix given by

$$\mathbf{\Lambda}_{ij} := \mathbf{A}_{ji} \frac{h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}|}{\sum_{\{\gamma_j \neq \emptyset\}} \langle h \rangle_{\gamma_j} \Pr(\gamma_j)} - \delta[i,j], \tag{10}$$

and $\mathbf{X}$ and $\mathbf{0}$ are column matrices of size $N \times 1$ with $\mathbf{X}_i = \lambda_i r_\infty(i)$ and $\mathbf{0}_i = 0$, and $\delta[i,j]$ is the Kronecker delta. To determine $s_\infty(i)$ one can either construct a similar expression to Eq. (9) or solve for $\mathbf{X}$ and then apply Eq. (8).

A unique solution[d] to Eq. (9) can be obtained upon introduction of the normalization condition $1 = \sum_i [r(i,t) + s(i,t)] = \sum_i [r_\infty(i) + s_\infty(i)]$ (details in Sect. A.6). For the general case where for each i the parameters ν_i, λ_i and h_i have different values, there is no simple closed form solution. However, for the simpler case $\nu_i = \nu$, the probability $\Pr(\gamma_i)$ simplifies to $\nu^{|\gamma_i|}(1 - \nu)^{k_i - |\gamma_i|}$, making the matrix elements $\mathbf{\Lambda}_{ij}$ take the form $\mathbf{A}_{ij} h_i/(k_j \langle h \rangle_{\Gamma_i}) - \delta[i,j]$, leading to

$$r_\infty^{(\nu)}(i) = \frac{\chi h_i \langle h \rangle_{\Gamma_i} k_i}{\lambda_i}, \tag{11}$$

$$s_\infty^{(\nu)}(i) = \frac{\chi h_i k_i}{1 - (1 - \nu)^{k_i}}, \tag{12}$$

$$\chi = \frac{1}{\sum_{i \in G} h_i \langle h \rangle_{\Gamma_i} k_i [\frac{1}{\lambda_i} + \frac{1}{\langle h \rangle_{\Gamma_i} [1 - (1-\nu)^{k_i}]}]}. \tag{13}$$

The expression for $s_\infty^{(\nu)}(i)$ can be rewritten by noting that the agent unemployed at i has the same probability ξ_i for any time step to find a job among its neighbors, given by $\xi_i = \langle h \rangle_{\Gamma_i} [1 - (1 - \nu)^{k_i}]$. This is because the likelihood that at least one of the Γ_i neighbors is open is $1 - (1 - \nu)^{k_i}$ for all time steps, and over configurations, the effective probability to be hired at any of them is $\langle h \rangle_{\Gamma_i}$. With this result

$$s_\infty^{(\nu)}(i) = \frac{\chi h_i \langle h \rangle_{\Gamma_i} k_i}{\xi_i}. \tag{14}$$

The rates λ_i and ξ_i play similar roles for employment and unemployment, respectively. An agent has a probability $\lambda_i (1 - \lambda_i)^{t-1}$ to be employed at i for t time steps, and similarly a probability $\xi_i (1 - \xi_i)^{t-1}$ to be unemployed t time steps at i. These are geometric distributions, for which the average times of employment (job tenure) and unemployment (spells) are, respectively, $1/\lambda_i$ and $1/\xi_i$. Fortunately, in data sets where the span of time spent unemployed is available (such as for the Mexican data we utilize here), ξ_i can be empirically estimated (together with k_i and λ_i), adding important practical value to this new parameter. As we explain below, ξ_i plays an important role when studying the unemployment consequences of our approach.

If the system has H agents, the steady state probabilities for the numbers of employed (L_i) and unemployed (U_i) agents at firm i can then be computed via

$$\Pr(L_i) = \binom{H}{L_i} \left[r_\infty(i) \right]^{L_i} \left[1 - r_\infty(i) \right]^{H-L_i} \tag{15}$$

and

$$\Pr(U_i) = \binom{H}{U_i} \left[s_\infty(i) \right]^{U_i} \left[1 - s_\infty(i) \right]^{H-U_i}. \tag{16}$$

These expressions are broadly useful because they are always valid in the steady state. Thus, even if $r_\infty(i)$ and $s_\infty(i)$ cannot be determined analytically but, say, numerically, their values can be used directly to determine $\Pr(L_i)$ and $\Pr(U_i)$.

One other important concept arises from this derivation: the quantity U_i can be interpreted as a *firm-specific unemployment*, which corresponds to the number of individuals that had i as their last employer but that are yet to find new employment. This notion is a powerful one, as it captures the essential nature of the network effect on mobility: if firm neighbors are not receiving new agents (probability $1 - v$), there is no place for the unemployed from i to go.

To illustrate some possible circumstances in which U_i can be useful we now discuss three scenarios. These represent examples of how the network introduces local effects to job markets, which cannot be captured in aggregate models, and can only be understood through firm-specific notions like those defined here. First, consider that there are regions of the LFN in which links are present because the connected firms hire similar individuals, and thus agents in one firm can more easily change jobs by going to the other firms in the same region of the network. This is a common situation and in numerous cases the regions of the network are rather cohesive (composed, say, of firms in the same or adjacent industrial sectors). This means that nodes in that region of the network are likely to react in similar ways to economic shocks. Therefore, if the firms in this region begin to lay off workers, they will quickly flood the local network neighborhood with job-seekers, most of which would have a hard time finding work. In addition to this, there would be nodes adjacent to the affected network region that would soon feel the effects of the employment shock by being flooded with new job seekers. In contrast, nodes in distant regions of the economy would feel a much more attenuated effect in a considerably longer time frame. In this shock scenario, the specific initial nodes affected, their unemployment values U_i, and the structure of the local network play significant roles.

There are other scenarios that can take place where U_i is informative. For example, in some cases links may be present due to differences in nodes rather than similarities (in two firms that react in opposite ways to an economic shock, for instance, one firm can become an alternative destinations for the workforce of the other firm). Knowing firm-specific unemployment for the neighboring nodes could be used to measure how anticorrelated their reactions are to economic shocks.

One more example is related to the steady state behavior of U_i for any given firm. Note that firms that have simultaneously large λ_i and h_i have a tendency to contribute large U_i to the network, and thus influence their neighboring nodes with a correspondingly large number of unemployed individuals seeking jobs, rapidly flooding their hiring capacities and then creating local unemployment in the network.

It is important to note that many individuals perform job searches while still employed. Therefore, to assume as we do that an agent must first separate from a job before looking for another job is an idiosyncratic choice. However, performing a job search while still employed is not fundamentally different from our approach since the individual still has to seek jobs with similar rules among firm neighbors, and thus the results of the Markov process are not qualitatively different, with the caveat that additional parameters are required to model the more nuanced case. Fortunately, as we see in our empirical analysis below, our current model assumptions seem to be effective in practice.

Note the versatility of our approach: since the equations are fully disaggregate, it is always possible to construct coarse-grained versions of the problem that can range from partially to fully aggregate. For instance, one known improvement over fully aggregate labor models are sectoral specialization models [12], which segment the labor market into submarkets each with its own set of shared employment mechanisms and parameters (also called matching technologies). In this context, one can picture the economy as constituted by an aggregation of several sectors (say, healthcare, technology, etc.) which have internal and external employment dynamics. In our model, this partial aggregation can be efficiently tackled by fully connecting all firms in a given sector and also adding links from each of these firms to all other firms in all other sectors of the economy. On the basis of the equations above, all firms of a given sector behave identically, effectively becoming a single *representative firm*. Representative firms connect to all other representative firms in the economy with appropriate weights (possibly adjusted on the basis of survey data such as the Job Openings and Labor Turnover Survey from the US Bureau of Labor Statistics). The result of applying our methods directly to the disaggregate description while introducing sectorized information or, alternatively, creating from scratch a reduced economy made of only representative firms are equivalent, and it is a matter of choice which approach to take, or even how to decide what firms are assigned to what sectors.

4 Empirical tests and results of the model

Armed with tools to determine the labor flow network as well as rules to model the behavior of agents, we now proceed to test the quality of our model. First, we focus on determining whether the system can be reliably modeled in the steady state. After that, we contrast the predictions emerging from the model with data from Finland and Mexico .[e] As we see below, the data and the model are consistent.
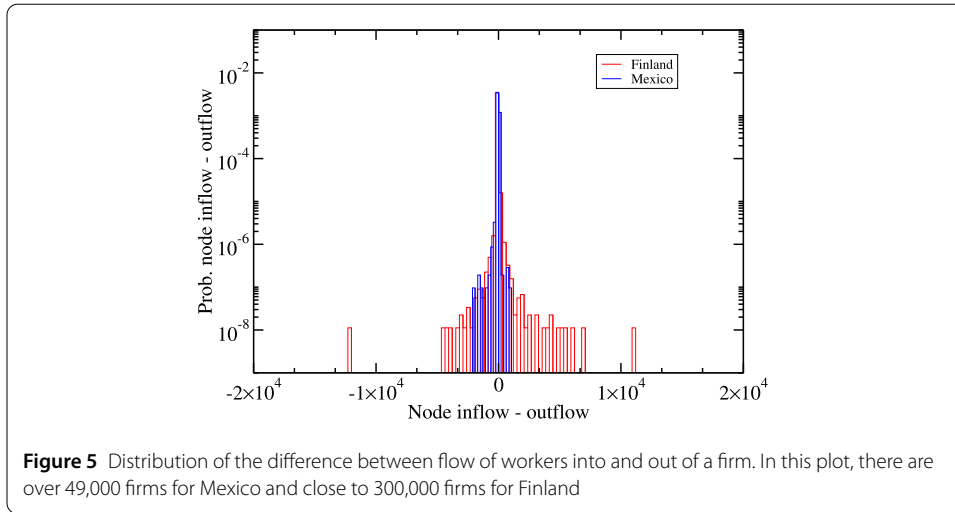
4.1 Testing the steady state assumption

In the previous section, we have constructed solutions for the steady state. In order to determine if such solutions are representative of typical economic situations, we check whether the number of individuals entering and exiting a node approximately match one another. In Fig. 5 we present the distribution of net flow at a node.

It is clear from our results that indeed firms typically operate around a steady state where the number of workers entering and leaving a firm approximately balance out. This provides empirical validity at the micro-level. For further support, in the Appendix we present several other tests that highlight the steady state nature of the system (see Sect. A.8).

4.2 Firm sizes

Data available to explore the dynamics of the firm-employee system do not contain information about $\{v_i\}_{i=1,\dots,N}$ or $\{h_i\}_{i=1,\dots,N}$. However, data is available to estimate the values of

Figure 5 Distribution of the difference between flow of workers into and out of a firm. In this plot, there are over 49,000 firms for Mexico and close to 300,000 firms for Finland

λ_i. For Finland, this is done by calculating the ratio of agents leaving a firm with respect to the size of the firm. A value of λ_i is created for every year of the sample, and then all the values for a given firm are averaged over those yearly samples.
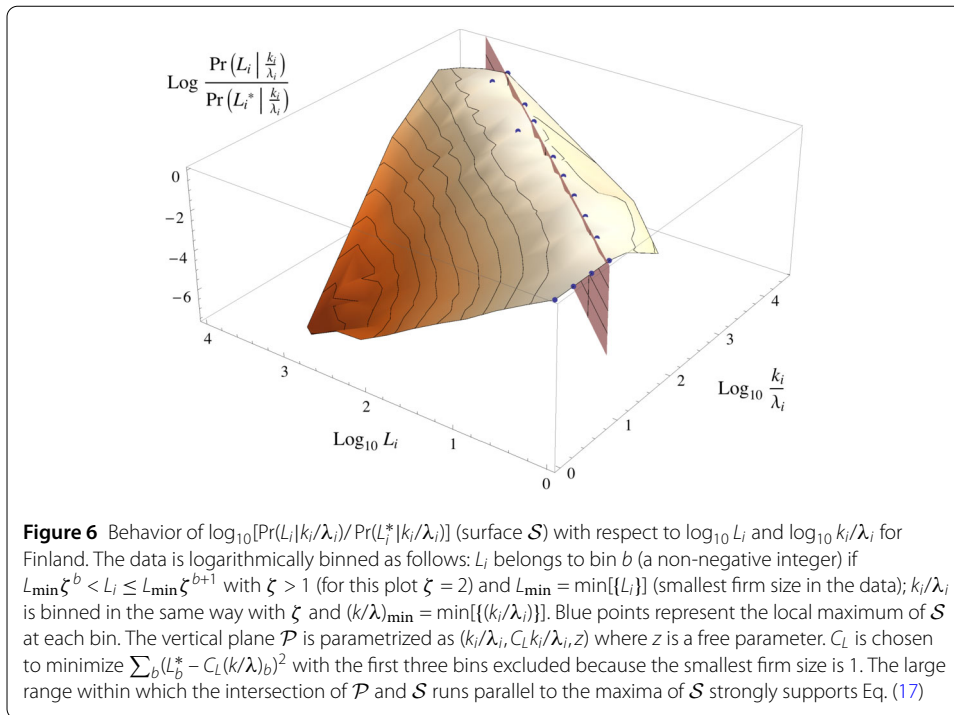
To deal with the absence of information for h_i, ν_i, we concentrate on comparing the data with the homogeneous $\nu_i = \nu$ version of our model in the hope that, if some of the main qualitative features of the system have been properly captured, we could find a reasonable level of agreement between data and prediction.

We now proceed to test the homogeneous model with $\nu_i = \nu$ (a constant) for all i. However, we retain the freedom for each firm i to have its own independent acceptance rate h_i in order to conform with experience (there are more selective and less selective firms in terms of hiring). Focusing first on the sizes of firms, we make use of Eqs. (11) and (15) to find that the most probable (or mode) firm size L_i^* is

$$L_i^* = \left\lfloor (H + 1) r_\infty^{(\nu)}(i) \right\rfloor \sim \frac{H \chi h_i \langle h \rangle_{\Gamma_i} k_i}{\lambda_i}. \tag{17}$$

This is a model prediction. One possible check for our model against the available data is to determine whether the measured L_i^* and k_i/λ_i relate as predicted by Eq. (17), assuming independence among the variables h_i, k_i, and λ_i. Furthermore, we can also study the distribution $\Pr(L_i|k_i/\lambda_i)$ to develop a broader picture.

To simultaneously learn about $\Pr(L_i|k_i/\lambda_i)$ and L_i^*, we use the Finnish data to generate Fig. 6, a 3-dimensional plot of $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ as a function of $\log_{10} L_i$ and $\log_{10}(k_i/\lambda_i)$, where each L_i is the size of a Finnish firm, k_i its degree based on repeated observed transitions, and λ_i its separation rate estimated over the years of data as explained above. Here, $\Pr(L_i^*|k_i/\lambda_i)$ is the probability associated with the conditional mode L_i^*. For a given k_i/λ_i, the logarithm of the ratio $\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)$ becomes 0 when $L_i = L_i^*$, and is < 0 for other values of L_i. To interpret the plot, we introduce a plane $\mathcal{P}$ parametrized as indicated in Fig. 6, normal to the base plane $\log_{10} L_i, \log_{10}(k_i/\lambda_i)$ and running parallel to its diagonal (which means it is a proportionality plane between L_i and k_i/λ_i). This normal plane also cuts the ratio $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ at or very close to 0, i.e., when $L_i = L_i^*$. Therefore, it means that L_i^* from the data is proportional to k_i/λ_i, supporting the prediction from Eq. (17). The correspondence we observe between the data and prediction

Figure 6 Behavior of $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ (surface $\mathcal{S}$) with respect to $\log_{10} L_i$ and $\log_{10} k_i/\lambda_i$ for Finland. The data is logarithmically binned as follows: L_i belongs to bin b (a non-negative integer) if $L_{\min}\zeta^b < L_i \leq L_{\min}\zeta^{b+1}$ with $\zeta > 1$ (for this plot $\zeta = 2$) and $L_{\min} = \min[\{L_i\}]$ (smallest firm size in the data); k_i/λ_i is binned in the same way with ζ and $(k/\lambda)_{\min} = \min[\{(k_i/\lambda_i)\}]$. Blue points represent the local maximum of $\mathcal{S}$ at each bin. The vertical plane $\mathcal{P}$ is parametrized as $(k_i/\lambda_i, C_L k_i/\lambda_i, z)$ where z is a free parameter. C_L is chosen to minimize $\sum_b (L_b^* - C_L(k/\lambda)_b)^2$ with the first three bins excluded because the smallest firm size is 1. The large range within which the intersection of $\mathcal{P}$ and $\mathcal{S}$ runs parallel to the maxima of $\mathcal{S}$ strongly supports Eq. (17)

also supports the assumption that, on average, the parameters h_i and $\langle h \rangle_{\Gamma_i}$ do not depend strongly on k_i or λ_i [15] (see Table 1 for a summary of the relation between parameters and results).

If indeed k_i/λ_i is strongly correlated to L_i as indicated by the results above, we can assume the relation
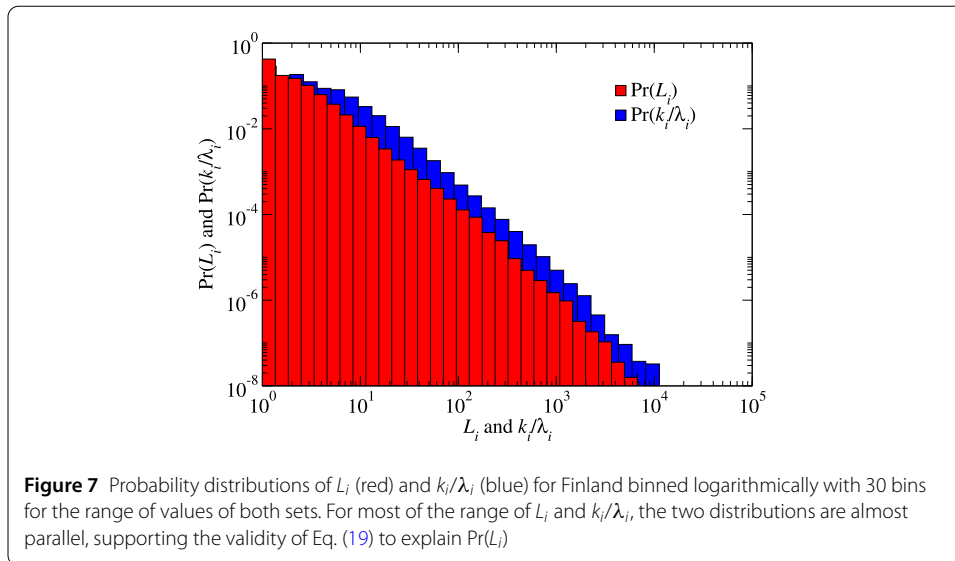
$$L_i \approx C_L \frac{k_i}{\lambda_i}, \tag{18}$$

where C_L is independent of k_i and λ_i (but still depends on the remaining model parameters $\nu, \{h_i\}, \chi$). Under this assumption, the distributions of both L_i and k_i/λ_i should be related by the change of variables theorem, which (written in the continuous limit) yields

$$\Pr(L_i)dL_i = \Pr(k_i/\lambda_i)d(k_i/\lambda_i). \tag{19}$$

In Fig. 7, we show the probability distributions of both L_i and k_i/λ_i which are close to parallel, and display a heavy tail, indeed supporting our assumption. It has been known for a long time that $\Pr(L_i)$ satisfies Zipf's law [20–23], which supports the notion that the probabilities in Fig. 7 follow a power law. Employing this functional form in Eq. (19) with decay exponent z, i.e., $\Pr(k_i/\lambda_i) \sim (k_i/\lambda_i)^{-z}$, we find

$$\Pr(L_i) \sim L_i^{-z}. \tag{20}$$

The value of z estimated from a least squares fit of the slope of $\log \Pr(k_i/\lambda_i)$ with respect to k_i/λ_i turns out to be $\approx 1.97 \pm 0.02$, consistent with the exponent of the decay of $\Pr(L_i)$ that we measure against our data (see Table 2). Note also that this exponent is consistent

Figure 7 Probability distributions of L_i (red) and k_i/λ_i (blue) for Finland binned logarithmically with 30 bins for the range of values of both sets. For most of the range of L_i and k_i/λ_i, the two distributions are almost parallel, supporting the validity of Eq. (19) to explain $\Pr(L_i)$

with the decay exponent close to -1 of the cumulative distributions of L_i observed for numerous countries [23].

This result offers a new interpretation for the origin of the power law distribution of firm sizes [20–22]. In our picture, the collection of employment affinities, and hence connectivity distribution, plays a dominant role (together with the separation rates) in the observed distribution of firm sizes. We should emphasize that this does not equate to a statement of causality (ultimate causes of employment affinity are structural variables such as geography, employee skills, etc.), but rather the realization that employment affinities are highly useful quantities with which to model because they possess a great power of synthesis about the system behavior.
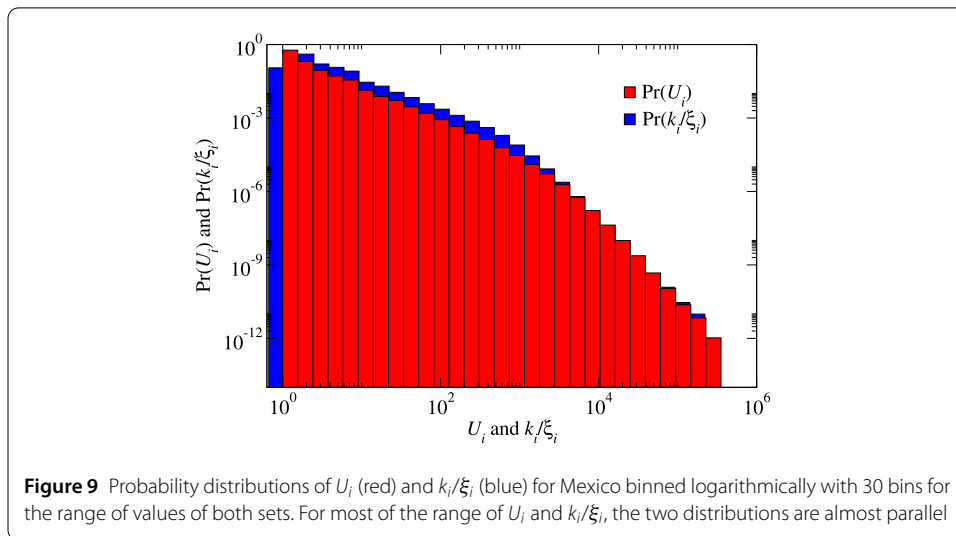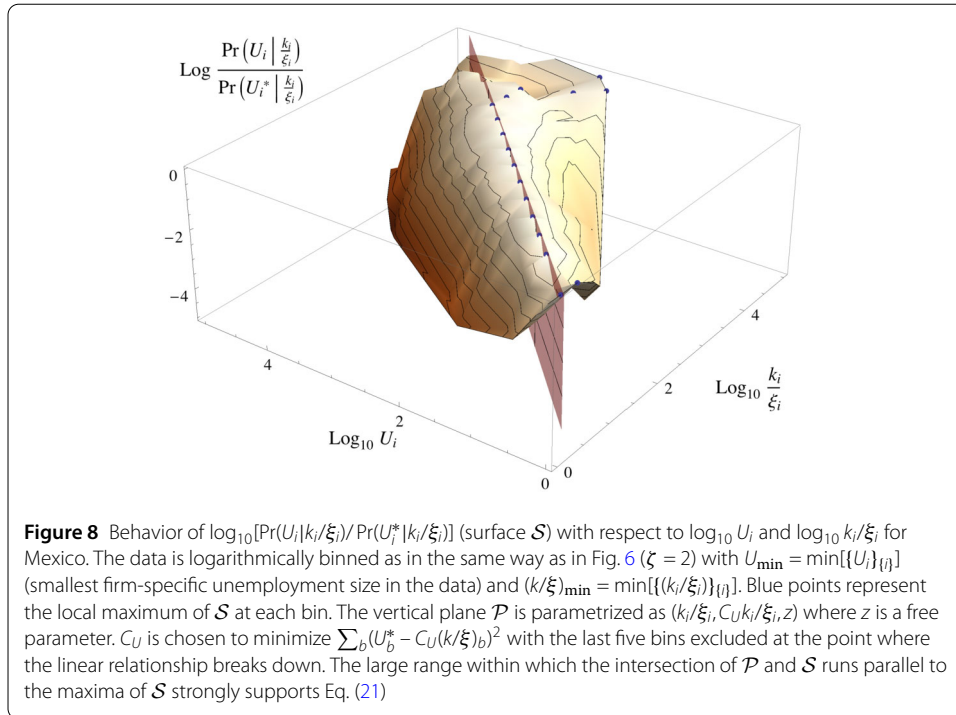
4.3 Firm-specific unemployment

The homogeneous model with $\nu_i = \nu$ also provides an estimate for our new concept of firm-specific unemployment. This quantity can be calculated in a similar way as L_i^*, and it is given by
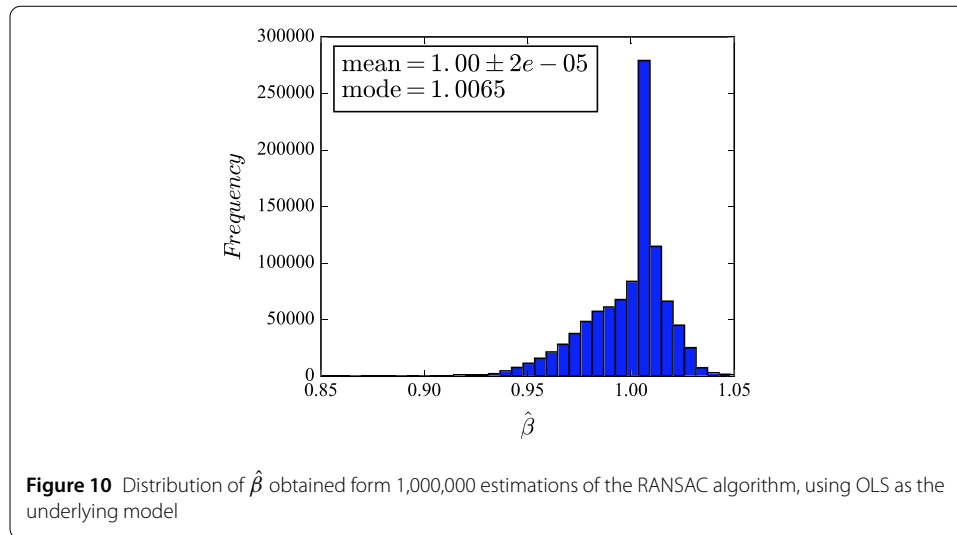
$$U_i^* \sim \frac{H \chi h_i \langle h \rangle_{\Gamma_i} k_i}{\xi_i}. \tag{21}$$

Although we do not have enough information in the Finnish dataset to determine the set $\{\xi_i\}$, we do for the Mexican dataset. From the latter, we determine the $\{\xi_i\}$ through maximum likelihood (see Sect. A.5). We test Eq. (21) in a similar way as Eq. (17), through a 3-dimensional plot of $\log_{10}[\Pr(U_i|k_i/\xi_i)/\Pr(U_i^*|k_i/\xi_i)]$ as a function of $\log_{10} U_i$ and $\log_{10}(k_i/\xi_i)$, where $U_i, k_i,$ and ξ_i are all determined empirically for Mexico. The results are shown in Fig. 8, and support the conclusion that $U^* \sim k_i/\xi_i$. For Mexico, we average U_i over the whole observation window of $D = 10{,}612$ days to obtain stable values for firm-specific unemployment.

To better understand firm-specific unemployment, we present in Fig. 9 its probability distribution. This is the first time this quantity is reported. Its importance revolves around the fact that firms which have a large contribution to unemployment may constitute a major problem for economies as a whole. In the same plot, we also display the probability

Figure 8 Behavior of $\log_{10}[\Pr(U_i|k_i/\xi_i)/\Pr(U_i^*|k_i/\xi_i)]$ (surface $\mathcal{S}$) with respect to $\log_{10} U_i$ and $\log_{10} k_i/\xi_i$ for Mexico. The data is logarithmically binned as in the same way as in Fig. 6 ($\zeta = 2$) with $U_{\mathbf{min}} = \min[\{U_i\}_{\{i\}}]$ (smallest firm-specific unemployment size in the data) and $(k/\xi)_{\mathbf{min}} = \min[\{(k_i/\xi_i)\}_{\{i\}}]$. Blue points represent the local maximum of $\mathcal{S}$ at each bin. The vertical plane $\mathcal{P}$ is parametrized as $(k_i/\xi_i, C_U k_i/\xi_i, z)$ where z is a free parameter. C_U is chosen to minimize $\sum_b (U_b^* - C_U(k/\xi)_b)^2$ with the last five bins excluded at the point where the linear relationship breaks down. The large range within which the intersection of $\mathcal{P}$ and $\mathcal{S}$ runs parallel to the maxima of $\mathcal{S}$ strongly supports Eq. (21)

Figure 9 Probability distributions of U_i (red) and k_i/ξ_i (blue) for Mexico binned logarithmically with 30 bins for the range of values of both sets. For most of the range of U_i and k_i/ξ_i, the two distributions are almost parallel

distribution of k_i/ξ_i. The parallel between the two distributions mirrors the situation with L_i and k_i/λ_i: Eq. (21), which connects U_i to k_i/ξ_i, allows explaining $\Pr(U_i)dU_i$ through a change of variables so that $\Pr(U_i)dU_i = \Pr(k_i/\xi_i)d(k_i/\xi_i)$. This plot also indicates the presence of a heavy-tail distribution for $\Pr(U_i)$. Some care must be taken in interpreting Fig. 9 due to the way in which the Mexican data was collected, focusing on uniformly sampling individuals rather than firms. This may play a role in the cross-over in slopes observed in the distributions of both U_i and k_i/ξ_i.

Figure 10 Distribution of $\hat{\beta}$ obtained form 1,000,000 estimations of the RANSAC algorithm, using OLS as the underlying model

4.4 Ratio of firm separation to waiting rates

An additional test can be performed on the basis of the symmetry between Eqs. (17) and (21). The ratio between these leads to

$$\frac{\langle L_i \rangle}{\langle U_i \rangle} = \frac{L_i^*}{U_i^*} = \alpha \left(\frac{\xi_i}{\lambda_i} \right)^{\beta} \tag{22}$$

with α and β equal to 1. To determine if the prediction is matched by the data, we applied the re-sample consensus algorithm (RANSAC) [37] (see Sect. A.10) with 10^6 estimations for the Mexican data. The results can be seen in Fig. 10. The average β is $1.00000 \pm 2 \times 10^{-5}$, while the most frequent is 1.0065. The average estimator α of the intercept is $1.13637 \pm 5 \times 10^{-5}$. These results are quite close to the theoretical prediction of (22).[f]

5 Discussion and conclusions

Detailed microdata, such as the one analyzed in this article, provides an opportunity to construct new, highly resolved models of macro level phenomena from micro level empirically justified mechanisms. In our particular case, our approach offers a picture of firms and employment that links them together in a precise way, opening the opportunity for an integrated theory of these two areas of research.

Our network picture of firms and employment offers the novel idea that the sizes of firms become encoded in the number of independent connections firms have with other firms. These connections, which reflect an economic affinity (low mobility barriers) relevant to employment transitions, synthesize the numerous possible structural variables (skills, geography, social contacts, etc.) that an agent is affected by when searching for employment, but because the connections are determined from the data (empirically calibrated), even those variables that may not be traditionally tracked are taken into account.

The ability of this data-driven approach to incorporate both known and unknown mechanisms in the firm-employee system makes our method less prone to idiosyncrasies associated with methods of modelling that require choosing a set of starting assumptions and then trying to model from that point on. For instance, if one had assumed that labor supply

was the main mechanism, the result would be that the model would not fit the data, and thus additional assumptions would have been needed to explain observation. This strategy would create a model that has to be adjusted *ad hoc*, is not parsimonious (due to the need for additional variables to control the adjustments), and is thus less tractable conceptually. Even if additional well-known mechanisms are incorporated into the modelling to try to achieve the adjustments, there is no guarantee that one can capture all relevant effects, producing the same incomplete and non-parsimonious modelling situation.

A new concept of firm-specific unemployment is also introduced here. From the standpoint of the theory of processes on graphs, it is a useful tool to account for a 'search' state of the agent, as one would see in queuing processes such as data routing on computer networks. In our particular case of employment and firm sizes, beyond its technical value, firm-specific unemployment introduces new economic notions about employment, relating to the relevance that specific firms along with their surroundings contribute to the overall unemployment rate.

The time scales that our model addresses (and their relation to real time scales), depend on whether the economy is steady enough that its behaviour between samples is not changing a great deal (see the discussion on the steady state in Sect. A.8) or if, in contrast, it is very dynamic. In the steady (or even in the slow dynamic) state, the model time scale and the real-world time scales basically match as the solutions to the model quickly adjust to the model parameters for that sampling period. When the dynamics are very rapid due to a fast economic shock, the time scales of the model need to be considered within the sampling period, a problematic situation since we would be unable to compare model results with data. In this regime, the time scale of the dynamics would be dictated by the first eigenvalue of the stochastic matrix of the Markov chain.

In the future, as our models improve and further data is gathered and analyzed, it may become possible to develop even more detailed models that could tackle more complex problems such as the formation of new firms and the construction of realistic shock scenarios, which are necessary to design real-time high resolution forecasting of employment flow. This task, which has not yet been possible, may be within our reach for the first time, with considerable potential for social policy design that is well grounded empirically and for which its effect can be forecast in great detail.

Appendix A
A.1 Data
We use two different datasets of employer-employee matched records. The first is the Finnish Longitudinal Employer-Employee Data (FLEED), which consists of an annual panel of employer-employee matched records of the universe of firms and employees in Finland. The panel was constructed by Statistics Finland from social security registries by recording the association between each worker and each firm (enterprise codes, not establishments), at the end of each calendar year. If a worker is not employed, it is not part of the corresponding cross-section. The result is a panel of 20 years (1988 to 2007) that tracks every firm and every employed individual at the end of each year (approximately 3×10^5 firms and 2×10^6 workers). From two consecutive years of this data, one can determine if an employee in one firm has moved to another, hence generating data for inter-firm job-to-job transitions. We have direct access to this data on transition, but not the entirety of FLEED.

Table 1 Parameters that are either used directly as inputs in our model, Eqs. (5) and (6), or used to determine or postulate those input parameters. The firm-size distribution (FSD) emerges for the conditions stated in the table, particularly that there are no interdependencies among the parameters for a node (one particularly problematic one would be that h_i and/or k_i were correlated, but our results do not support this)

Parameter	Empirical	Model	Model conditions for FSD
λ_i	✓	✓	–
ξ_i	✓	✓ (function of h_i, v_i)	–
h_i	–	✓	$0 < h_i \leq 1$, no parameter interdependencies
v_i	–	✓	$0 < v_i \leq 1$, no parameter interdependencies
f_{ij}	✓	–	–
a_{ij}	–	✓ (determined from f_{ij})	–

Table 2 Quantities predicted by the model against measured counterparts

Quantity	Prediction	Measurement
z [Eq. (20)]	-1.97 ± 0.02 [from $\Pr(k_i/\lambda_i)$]	-1.97 ± 0.03 [from $\Pr(L_i)$]
α [Eq. (22)]	1	$1.13637 \pm 5 \times 10^{-5}$
β [Eq. (22)]	1	$1.00000 \pm 2 \times 10^{-5}$

Unemployment periods cannot be determined from FLEED. For this we use a dataset from Mexico consisting of employer-employee matched records with daily resolution. The data was obtained by sampling raw social security records from the Mexican Social Security Institute. Approximately 4×10^5 individuals who were active between 1989 and 2008 were randomly selected and their entire employment history was extracted (hence, covering dates prior to 1989). This procedure generates a dataset with nearly 2×10^5 firms. The records contain information about the exact date in which a person became hired/separated by/from a firm. Therefore, it is possible to identify unemployment spells, duration of each spell, and associations between job seekers and their last employer.

As a supplementary dataset, useful for determining sizes of firms and separation rates, we use Statistics Finland's Business Register, constructed from administrative data from the Tax Administration, and from direct inquiries from Statistics Finland to business with more than 20 employees. This data provides firm sizes and profits from different sources.

A.2 Summary of empirical and model parameters. Model predictions
To facilitate the presentation as well as provide a summary of the role of the parameters of the model, we present in this section Table 1 with all the parameters that bear relevance to the inputs of the model. We also indicate the conditions under which these parameters are consistent with the empirical observations we attempt to reproduce in this work, namely the firm-size distribution (FSD).

In Table 2, we provide a summary of the quantities predicted by the model and those that are measured empirically and matched against the model.

A.3 Testing persistence in the Finnish dataset
In this section we explain further details, particularly regarding the null models, about the effectiveness of the threshold criterion in determining which node pairs should be connected using a link on the basis of the flow of individuals between the firms.

A.3.1 Null model and p-values for threshold criterion

To address the construction of a *p*-value for the analysis of persistence carried out in Sect. 2.1, we can define the null model mathematically. For this purpose, suppose that among the nodes $\mathcal{N}^*(t, \Delta t, \mathcal{W})$, the null model is defined by the fact that the flows captured in $\mathcal{E}^*(t, \Delta t, \mathcal{W})$ and those captured in $\bar{\mathcal{E}}^*(t, \Delta t, 1)$ overlap only as a consequence of random chance. Concretely, every flow in $\mathcal{E}^*(t, \Delta t, \mathcal{W})$ that also belongs to $\bar{\mathcal{E}}^*(t, \Delta t, 1)$ is considered a successful random trial, which takes away a success state (sampling without replacement). The success states are the flows in $\bar{\mathcal{E}}^*(t, \Delta t, 1)$. The overall population, i.e., places where the flows $\bar{\mathcal{E}}^*(t, \Delta t, 1)$ can be placed, consists of all the unique pairs of nodes among $\mathcal{N}^*(t, \Delta t, \mathcal{W})$. Therefore, the likelihood that there are $|\mathcal{E}^*(t, \Delta t, \mathcal{W}) \cap \bar{\mathcal{E}}^*(t, \Delta t, 1)|$ successful trials is given by the hypergeometric distribution

$$\Pr\left(\left|\mathcal{E}^* \cap \bar{\mathcal{E}}^*\right|\right) = \frac{\binom{|\mathcal{E}^*|}{|\mathcal{E}^* \cap \bar{\mathcal{E}}^*|}\binom{\binom{|\mathcal{N}^*|}{2}-|\mathcal{E}^*|}{|\bar{\mathcal{E}}^*|-|\mathcal{E}^* \cap \bar{\mathcal{E}}^*|}}{\binom{\binom{|\mathcal{N}^*|}{2}}{|\bar{\mathcal{E}}^*|}}, \tag{23}$$

where, for brevity, we use the shortened notation $\mathcal{E}^*_{\mathcal{W}}, \bar{\mathcal{E}}^*_1$, and $\mathcal{N}^*_{\mathcal{W}}$. Therefore, the expectation value $\langle|\mathcal{E}^*_{\mathcal{W}} \cap \bar{\mathcal{E}}^*_1|\rangle$ for $|\mathcal{E}^*_{\mathcal{W}} \cap \bar{\mathcal{E}}^*_1|$ is given by

$$\left\langle\left|\mathcal{E}^*_{\mathcal{W}} \cap \bar{\mathcal{E}}^*_1\right|\right\rangle = \frac{|\bar{\mathcal{E}}^*_1||\mathcal{E}^*_{\mathcal{W}}|}{\binom{|\mathcal{N}^*_{\mathcal{W}}|}{2}}, \tag{24}$$

where $|\mathcal{E}^*_{\mathcal{W}}|/\binom{|\mathcal{N}^*_{\mathcal{W}}|}{2}$ is the probability of picking a pair of nodes among $\mathcal{N}^*_{\mathcal{W}}$ between which there is flow that belongs to $\mathcal{E}^*_{\mathcal{W}}$; if $\Pr(|\mathcal{E}^*_{\mathcal{W}} \cap \bar{\mathcal{E}}^*_1|)$ were given by a binomial instead, the expectation value would be the same (this is relevant below). Rewriting the last expression somewhat, we find

$$\frac{\langle|\mathcal{E}^*_{\mathcal{W}} \cap \bar{\mathcal{E}}^*_1|\rangle}{|\mathcal{E}^*_{\mathcal{W}}|} = \frac{|\bar{\mathcal{E}}^*_1|}{\binom{|\mathcal{N}^*_{\mathcal{W}}|}{2}} = \wp, \tag{25}$$

where the second equality comes from the definition of $\wp$ in Eq. (2) above. Note also that the left hand side is the expectation value for $\wp_{\mathcal{W}}$ in the null model. In other words, Eq. (25) is a proof of our statement that, in the random model, the expected value for $\wp_{\mathcal{W}}$ should correspond to $\wp$. However, the observed $\wp_{\mathcal{W}}$ are much larger than $\wp$, supporting the use of $\mathcal{W}$ as a selection criterion for links. To estimate a *p*-value using the hypergeometric distribution is difficult because the values of $\mathcal{E}^*_{\mathcal{W}}, \bar{\mathcal{E}}^*_1$, and $\mathcal{N}^*_{\mathcal{W}}$ are quite large (see the table below). Therefore, we estimate the *p*-values using a normal distribution which can, in turn, be explained from a binomial distribution approximation to the hypergeomatric distribution, well justified in our case given that $\binom{|\mathcal{N}^*_{\mathcal{W}}|}{2} \gg |\mathcal{E}^*_{\mathcal{W}}|, |\bar{\mathcal{E}}^*_1|$. In the binomial approximation, in order to maintain the same expectation value, the success probability ρ_s is given by

$$\rho_s = \frac{|\mathcal{E}^*_{\mathcal{W}}|}{\binom{|\mathcal{N}^*_{\mathcal{W}}|}{2}}, \tag{26}$$

which says that to pick a node pair where there was flow during $\mathcal{T}_<$, the chances are proportional to the number of node pairs $|\mathcal{E}^*_{\mathcal{W}}|$ in that time period. In the normal approximation

to this binomial distribution, the mean and standard deviation are given by

$$\mu = \left|\bar{\mathcal{E}}_1^*\right|\rho_s, \mathcal{Q} = \sqrt{\left|\bar{\mathcal{E}}_1^*\right|\rho_s(1-\rho_s)} \tag{27}$$

since $|\bar{\mathcal{E}}_1^*|$ is the number of trials. For this approximation, the p-value is then given by the integral

$$p\text{-value of } \left|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*\right| \approx \frac{1}{\sqrt{2\pi \mathcal{Q}^2}} \int_{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|}^{\infty} e^{-\frac{(\phi-\mu)^2}{2\mathcal{Q}^2}}\, d\phi. \tag{28}$$

On the basis of the values in the table below, the magnitude of ϕ is very large in comparison to μ and therefore, one can approximate the integral via an asymptotic expansion of first order derived by integration by parts, giving estimates for the p-value of

$$p\text{-value of } \left|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*\right| \approx \frac{1}{\sqrt{2\pi}} \frac{\mathcal{Q}}{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu} e^{-\frac{(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|-\mu)^2}{2\mathcal{Q}^2}}. \tag{29}$$

Subsequent terms in the expansion are also dominated by the exponential term and therefore, it is reasonable to truncate the expansion at first order. The values of the exponent are large enough that it is better to express these results under a logarithm, producing

$$\ln\left[p\text{-value of } \left|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*\right|\right]$$

$$\approx -\frac{(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu)^2}{2\mathcal{Q}^2} + \ln\left[\frac{1}{\sqrt{2\pi}} \frac{\mathcal{Q}}{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu}\right]$$

$$\approx -\frac{(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu)^2}{2\mathcal{Q}^2} - \ln\left[\frac{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu}{\sqrt{2}\mathcal{Q}}\right] - \ln(2\sqrt{\pi})$$

$$\approx -\frac{(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \mu)^2}{2\mathcal{Q}^2} = -\frac{(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*| - \langle|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|\rangle)^2}{2\mathcal{Q}^2}, \tag{30}$$

where the last equality uses the fact that $\mu = |\bar{\mathcal{E}}_1^*||\mathcal{E}_{\mathcal{W}}^*|/\binom{|\mathcal{N}_{\mathcal{W}}^*|}{2} = \langle|\mathcal{E}_1^* \cap \bar{\mathcal{E}}_1^*|\rangle$, and drops the logarithmic term and the constant since they are much smaller in magnitude than the quadratic term.

The p-value estimates are contained in Table 3. The order of magnitude of these results is overwhelmingly below the usual significance threshold of 10^{-3}. To illustrate this with one of the combinations of values below ($t = 1989, \Delta t = 2, \mathcal{W} = 1$), note that $\rho_s = 2.2 \times 10^{-4}$, $\mu = |\bar{\mathcal{E}}_1^*|\rho_s = 8.78$, and $\mathcal{Q} = \sqrt{|\bar{\mathcal{E}}_1^*|\rho_s(1-\rho_s)} = 2.96$. Therefore, our estimate produces $\ln[p\text{-value}] \approx -4.64 \times 10^6 \ll \ln 10^{-3}$. For reference, $\ln 10^{-3} \approx -6.91$ where 10^{-3} comes from a p-value $= 10^{-3}$. All other results in the table show similar behavior, orders of magnitude removed from the 10^{-3} significance threshold. This provides convincing evidence for the validity of our method. The results of our tests for several combinations of years, time windows Δt, and values of $\mathcal{W}$ is shown in Table 3.

A.3.2 Null model and p-values for threshold criterion for heterogeneous firms

The p-values for the null model addressing the heterogeneity are calculated from the Monte Carlo simulations. As explained in the main text, we generate random realizations

Table 3 Table for test of threshold method for the cases that ignore or take into account the firm heterogeneity (HF). The table only includes $\Delta t = 2$, as other values of Δt have only a minor effect on the probability densities $\wp, \wp^{(\mathrm{HF})}$, and $\wp_{\mathcal{W}}$. We show a selection of years for each of the three values of $\mathcal{W} = 1, 2, 3$. Note that the p-values are exceedingly small for both tests

t	$\mathcal{W}$	$|\mathcal{N}_{\mathcal{W}}^*|$	$|\mathcal{E}_{\mathcal{W}}^*|$	$|\bar{\mathcal{E}}_1^*|$	$\mu = \langle|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|\rangle$ (ln[p-value])	$\mu_{\mathrm{HF}} = \langle|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|\rangle$ (ln[p-value])	$|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|$
1990	1	20,181	49,855	25,213	6.17 (-3.16×10^6)	1144.30 (-1.78×10^4)	6253
1995	1	22,707	33,398	39,424	5.11 (-3.50×10^6)	975.66 (-2.01×10^4)	5986
2000	1	32,913	68,225	64,381	8.11 (-6.83×10^6)	1793.58 (-3.56×10^4)	10,532
2005	1	40,037	72,123	88,335	7.95 (-8.64×10^6)	2278.98 (-3.15×10^4)	11,727
1990	2	6505	7584	12,986	4.66 (-7.08×10^5)	516.73 (-9.23×10^3)	2571
1995	2	5200	4470	16,817	5.56 (-3.51×10^5)	398.19 (-6.93×10^3)	1981
2000	2	7888	8462	27,800	7.56 (-8.92×10^5)	761.25 (-1.40×10^4)	3680
2005	2	8584	7630	38,601	8.00 (-7.95×10^5)	830.87 (-1.09×10^4)	3573
1990	3	3488	3202	8494	4.47 (-2.61×10^5)	340.20 (-5.18×10^3)	1532
1995	3	2631	1853	10,403	5.57 (-9.83×10^4)	219.36 (-4.84×10^3)	1052
2000	3	3716	3263	17,121	8.09 (-2.37×10^5)	448.39 (-7.86×10^3)	1967
2005	3	4388	2826	25,735	7.56 (-2.10×10^5)	474.05 (-5.33×10^3)	1789

$\bar{\mathcal{S}}_1^*$ of the flows $\bar{\mathcal{E}}_1^*$ (for $\mathcal{T}_>$) that respect the total in- and out-flows of each of the nodes in $\mathcal{N}_{\mathcal{W}}^*$. We then determine for each $\bar{\mathcal{S}}_1^*$ its overlap $\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*$ with the flows in the period $\mathcal{T}_<$. The equivalent quantity to $\wp$ of the previous test is $\wp^{(\mathrm{HF})} = \frac{\langle|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|\rangle_{(\mathrm{HF})}}{|\mathcal{E}_{\mathcal{W}}^*|}$ (defined in Eq. (4)) or, alternatively we can calculate the distribution $\mathrm{Pr}(\frac{|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|}{|\mathcal{E}_{\mathcal{W}}^*|})$ from the values of $|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|/|\mathcal{E}_{\mathcal{W}}^*|$ of each realization. Both $\wp^{(\mathrm{HF})}$ or the distribution over realizations can be compared against $\wp_{\mathcal{W}}$, but we require the distribution to determine a p-value. Our results indicate that a Gaussian fits well the distribution $\mathrm{Pr}(\wp^{(\mathrm{HF})})$.

Since both $\wp$ and $\wp^{(\mathrm{HF})}$ are defined with the $|\mathcal{E}_{\mathcal{W}}^*|$ in the denominator, we can determine p-values equally well by comparing $\mathrm{Pr}(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|)$ and $|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{E}}_1^*|$. This can be seen in a general way using $\phi = \tilde{\phi}/\alpha$ and $\mu = \tilde{\mu}/\alpha$, which lead to the identity
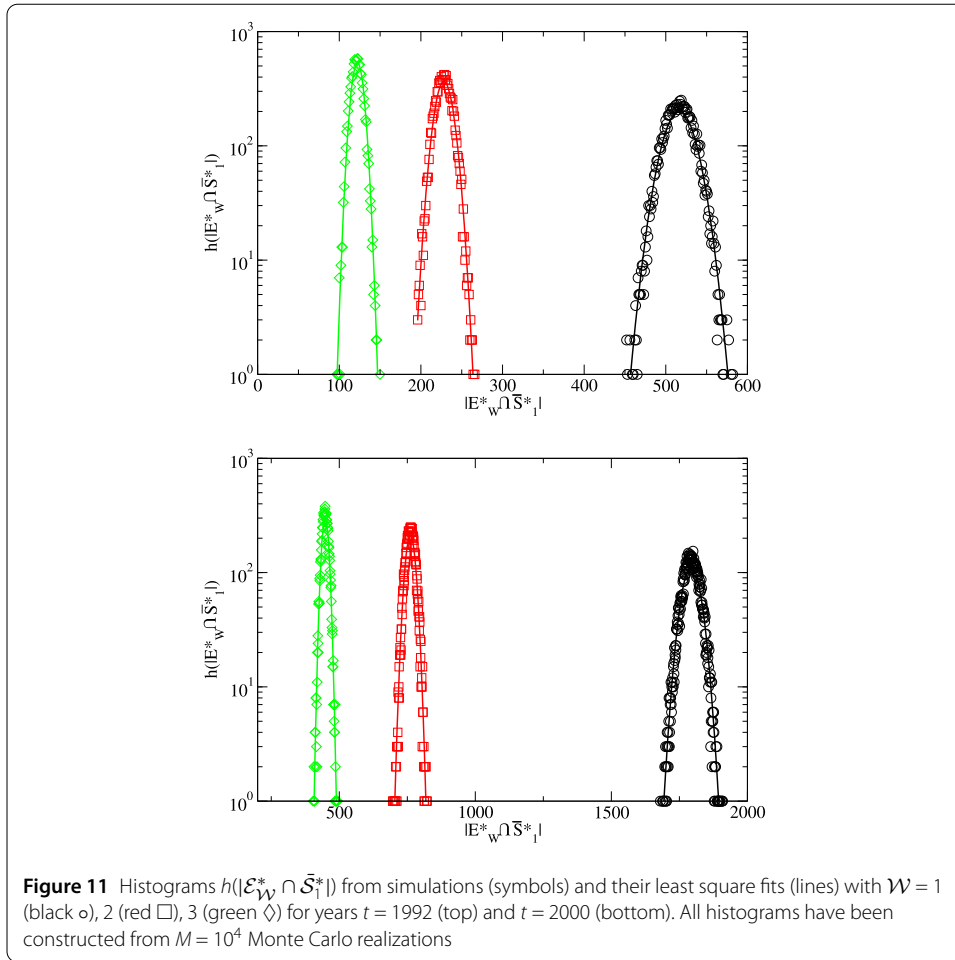
$$\frac{1}{\sqrt{2\pi}\sigma} \int_{\phi_o}^{\infty} e^{-\frac{(\phi-\mu)^2}{2\sigma^2}} \, d\phi = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{\phi_o/\alpha}^{\infty} e^{-\frac{(\tilde{\phi}-\tilde{\mu})^2}{2\tilde{\sigma}^2}} \, d\tilde{\phi}, \tag{31}$$

where σ is the standard deviation for x, and its transformed version $\sigma\alpha = \tilde{\sigma}$ provides the standard deviation for $\tilde{\phi}$. This transformation also highlights that, in terms of functional form, if $\mathrm{Pr}(\wp^{(\mathrm{HF})})$ is Gaussian, then so is $\mathrm{Pr}(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|)$.

In Fig. 11(a) we present histograms $h(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|)$ for $t = 1992$ and in Fig. 11(b) histograms for $t = 2000$, both for $\mathcal{W} = 1, 2, 3$, as well as their regressions (described next). Since we do not have a theoretical way to specify $\langle|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|\rangle_{(\mathrm{HF})}$ or its distribution $\mathrm{Pr}(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|)$, we resort to regression to determine the parameters of the histogram and subsequently estimate p-values using Eq. (30) (results in Table 3). With $\phi = |\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|$, we use

$$h(\phi) = h_o e^{-\frac{(\phi-\mu_{\mathrm{HF}})^2}{2\sigma_{\mathrm{HF}}^2}}, \tag{32}$$

where μ_{HF} and σ_{HF} emerge from the regression. These are the two parameters we care about, as the normalization constant of the Gaussian is determined by σ_{HF}. The quality of the fit is clear from Fig. 11.

Figure 11 Histograms $h(|\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*|)$ from simulations (symbols) and their least square fits (lines) with $\mathcal{W} = 1$ (black ○), 2 (red □), 3 (green ◇) for years $t = 1992$ (top) and $t = 2000$ (bottom). All histograms have been constructed from $M = 10^4$ Monte Carlo realizations

A.3.3 Overall flows captured by threshold criterion

The final test we carry out concerns how the threshold criterion performs with regards to accounting for the overall flow in the network.

Specifically, we define the total flow

$$\bar{\mathcal{F}}_{\cdot}^*\big(t, \Delta t | \bar{\mathcal{E}}^*(t, \Delta t, 1)\big) = \sum_{(i,j) \in \bar{\mathcal{E}}^*(t, \Delta t, 1)} \bar{\mathcal{F}}_{i,j}^*(t, \Delta t), \tag{33}$$

that is, the *total* number of individuals changing firms over the pairs of nodes in $\bar{\mathcal{E}}^*(t, \Delta t, 1)$. As above, we use the simplified notation $\bar{\mathcal{F}}_{\cdot}^*(\bar{\mathcal{E}}_1^*)$ to mean $\bar{\mathcal{F}}_{\cdot}^*(t, \Delta t | \bar{\mathcal{E}}^*(t, \Delta t, 1))$ as typically there is no confusion about the time t and time window Δt we are concentrating on. Note that $\bar{\mathcal{F}}_{\cdot}^*(\bar{\mathcal{E}}_1^*)$ counts all the repeated transitions between node pairs, i.e., all the flows happening during $\mathcal{T}_>$.

Our test consists of comparing $\bar{\mathcal{F}}_{\cdot}^*(\bar{\mathcal{E}}_1^*)$ to the average of the quantity

$$\bar{\mathcal{F}}_{\cdot}^*\big(t, \Delta t | \mathcal{E}^*(t, \Delta t, \mathcal{W}) \cap \bar{\mathcal{E}}^*(t, \Delta t, 1)\big) = \sum_{(i,j) \in \mathcal{E}^*(t, \Delta t, \mathcal{W}) \cap \bar{\mathcal{E}}^*(t, \Delta t, 1)} \bar{\mathcal{F}}_{i,j}^*(t, \Delta t), \tag{34}$$

which is the number of individuals that change jobs in the time period $\mathcal{T}_>$ *exclusively* on those node pairs that also had flows $\geq \mathcal{W}$ in the time period $\mathcal{T}_<$. We denote this flow as

$\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{E}}_1^*)$. The main purpose of this test is to show that a considerable percentage of future flow is captured using the threshold criterion, much more than the flow predicted by the null model for heterogeneous firm sizes presented above in Sects. 2.1 and A.3.2.

The overall flow during $\mathcal{T}_>$ associated with a given Monte Carlo realization of the null model for heterogeneous firm sizes is denoted by $\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*)$ where $\bar{\mathcal{S}}_1^*$ is the set of flows that emerges from one random rewiring that preserves in- and out-flows for each and every firm. As we discussed in Sect. A.3.2, flows $\mathcal{T}_>$ are reliably predicted by flows in the interval $\mathcal{T}_<$. Thus, we expect that typically $\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{E}}_1^*)$ is greater or even much greater than any typical $\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*)$. To confirm this is the case, we generate the histogram $h(\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*))$ of the overall flows predicted by the null model in order to compute $\Pr(\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*))$, its average $\langle \bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*) \rangle_{\text{HF}}$, and the p-value for the actual measured flow $\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{E}}_1^*)$. In addition, following the same conceptual framework as for the previous two tests, we define the fractions

$$\theta_\mathcal{W}(t, \Delta t) = \frac{\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{E}}_1^*)}{\bar{\mathcal{F}}_.^*(\bar{\mathcal{E}}_1^*)} \tag{35}$$

and

$$\theta^{(\text{HF})}(t, \Delta t) = \langle \theta(t, \Delta t) \rangle_{\text{HF}} = \frac{\langle \bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*) \rangle_{\text{HF}}}{\bar{\mathcal{F}}_.^*(\bar{\mathcal{E}}_1^*)}, \tag{36}$$

and the excess probability

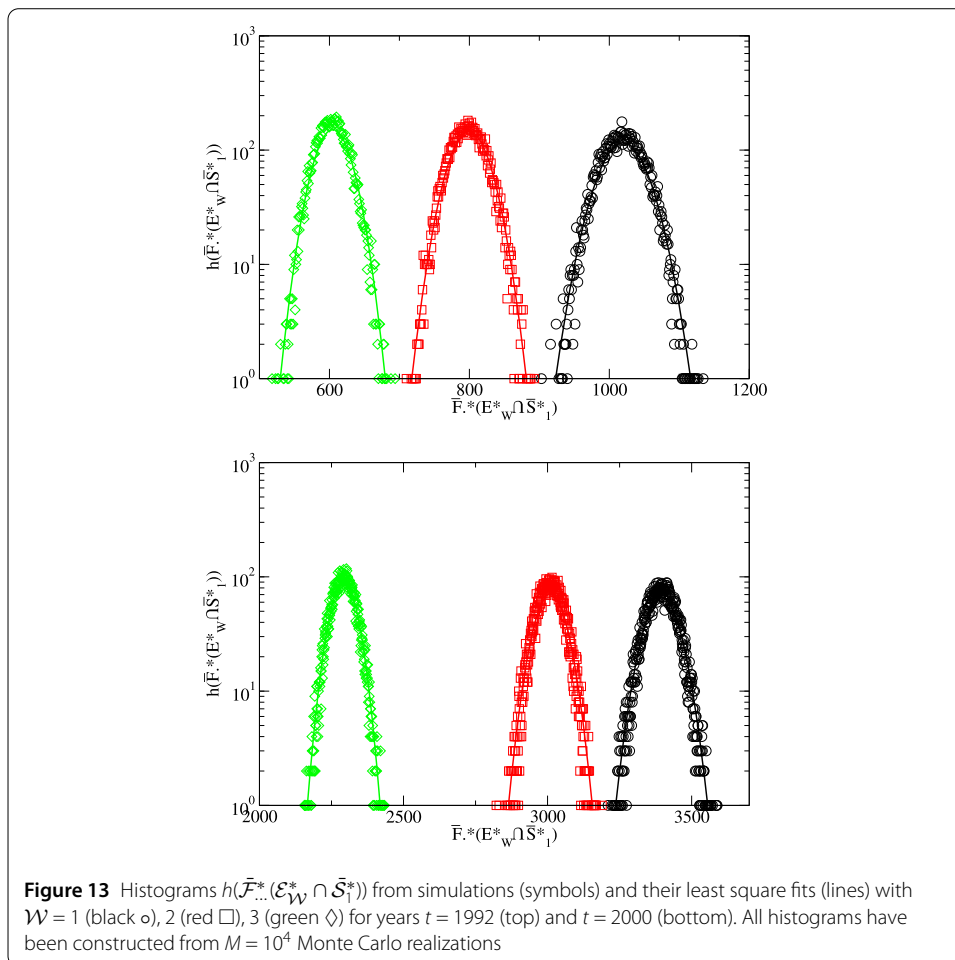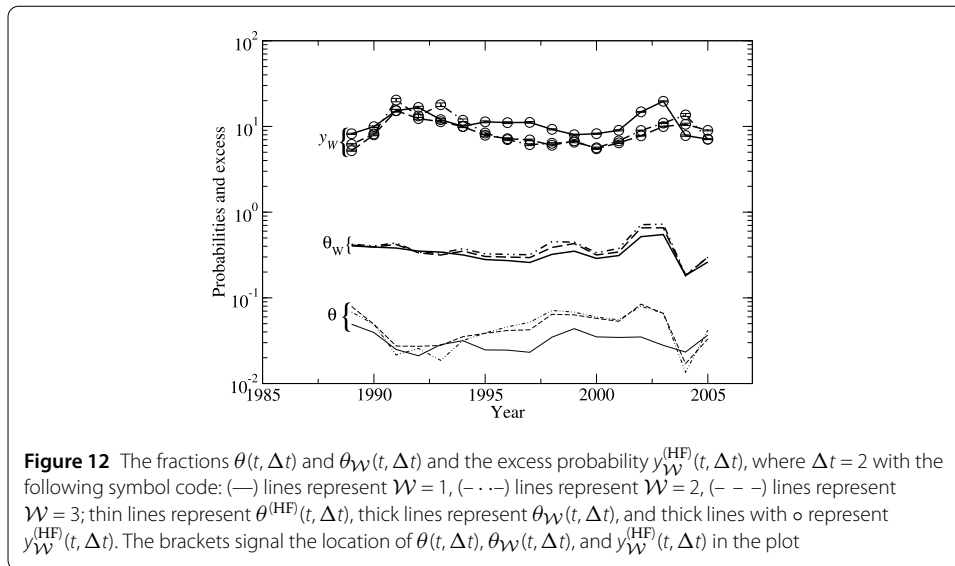$$y_\mathcal{W} = \frac{\theta_\mathcal{W}(t, \Delta t)}{\theta^{(\text{HF})}(t, \Delta t)}. \tag{37}$$

Figure 12 corresponds to $\theta_\mathcal{W}, \theta^{(\text{HF})}$, and $y_\mathcal{W}$. In consistent fashion with the results for $\wp^{(\text{HF})}, \wp_\mathcal{W}$, and $x_\mathcal{W}$, the values of $\theta_\mathcal{W}$ are roughly an order of magnitude larger than those for $\theta^{(\text{HF})}$, measured by $y_\mathcal{W}$. The values for $\theta_\mathcal{W}$ in particular are largely in a range that spans between 30% and 50% and reaches as high as just over 70%.

The calculation of p-values follows the same approach as the two previous tests. Figure 13 shows the histograms $h(\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*))$ for $\mathcal{W} = 1, 2, 3$ and the years (a) $t = 1992$ and (b) $t = 2000$, along with their least square fit (solid lines). The histograms were create with $M = 10^4$ rewirings of each individual yearly network. From the fits we estimate $\Pr(\bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*))$, its average $\langle \bar{\mathcal{F}}_.^*(\mathcal{E}_\mathcal{W}^* \cap \bar{\mathcal{S}}_1^*) \rangle_{\text{HF}}$ and standard deviation, which we then use in estimating the p-value of each set of years and thresholds. The results of this analysis are shown in Table 4.

A.4 Topological properties of labor flow network

The use of the criteria developed in Sects. 2.1 and A.3 leads to a labor flow network with certain topological properties that are captured in the distributions of degree, node strength, and link weights. These distributions are found in Figs. 14, 15, and 16 for a range of values of $\mathcal{W}$. A satisfactory feature of these distributions is that their functional forms remain stable even as $\mathcal{W}$ increases.
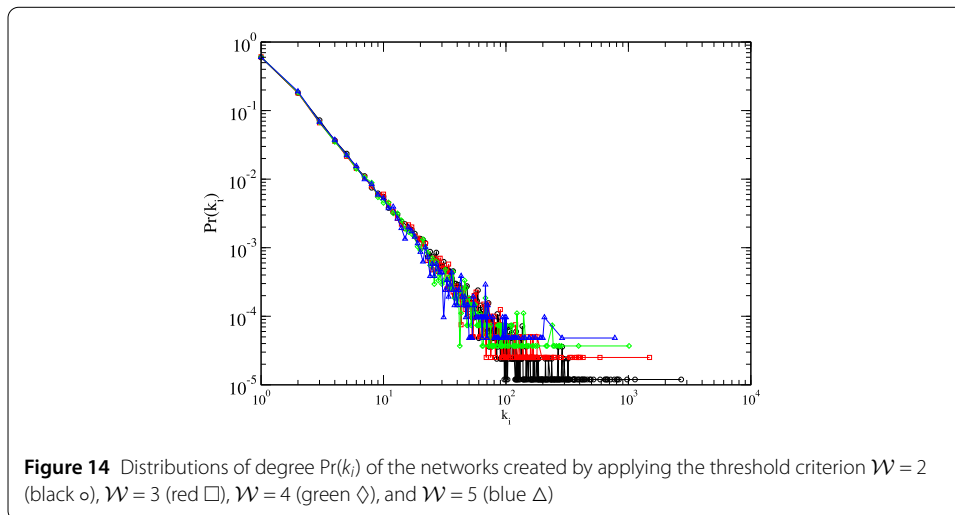
An important observation that it worth highlighting is that, while k_i and τ_i are clearly related, one is not predictive of the other. In other words, as has been found in prior research on real weighted networks compared to weighted network models, a full description of such networks typically requires knowledge of all the variables describing a node i,

Figure 12 The fractions $\theta(t, \Delta t)$ and $\theta_{\mathcal{W}}(t, \Delta t)$ and the excess probability $y_{\mathcal{W}}^{(HF)}(t, \Delta t)$, where $\Delta t = 2$ with the following symbol code: (—) lines represent $\mathcal{W} = 1$, (– · ·–) lines represent $\mathcal{W} = 2$, (– – –) lines represent $\mathcal{W} = 3$; thin lines represent $\theta^{(HF)}(t, \Delta t)$, thick lines represent $\theta_{\mathcal{W}}(t, \Delta t)$, and thick lines with ○ represent $y_{\mathcal{W}}^{(HF)}(t, \Delta t)$. The brackets signal the location of $\theta(t, \Delta t)$, $\theta_{\mathcal{W}}(t, \Delta t)$, and $y_{\mathcal{W}}^{(HF)}(t, \Delta t)$ in the plot

Figure 13 Histograms $h(\bar{\mathcal{F}}_{...}^* (\mathcal{E}_{\mathcal{W}}^* \cap \bar{\mathcal{S}}_1^*))$ from simulations (symbols) and their least square fits (lines) with $\mathcal{W} = 1$ (black ○), 2 (red □), 3 (green ◇) for years $t = 1992$ (top) and $t = 2000$ (bottom). All histograms have been constructed from $M = 10^4$ Monte Carlo realizations

including k_i, τ_i and the distribution of the link weights F_{ij} [38]. The relevance of this point stems from the fact that it confirms that the labor flow network is not simply a direct effect of the sizes of firms present in the economy: the observed k_i is not trivially predicted as the

Table 4 Table for test of overall flow carried by node pairs predicted by threshold test. The table only includes $\Delta t = 2$, as other values of Δt have only a minor effect on the probability densities $\theta^{(\mathrm{HF})}$ and $\theta_{\mathcal{W}}$. We show a selection of years for each of the three values of $\mathcal{W} = 1, 2, 3$. Note that the *p*-values are exceedingly small
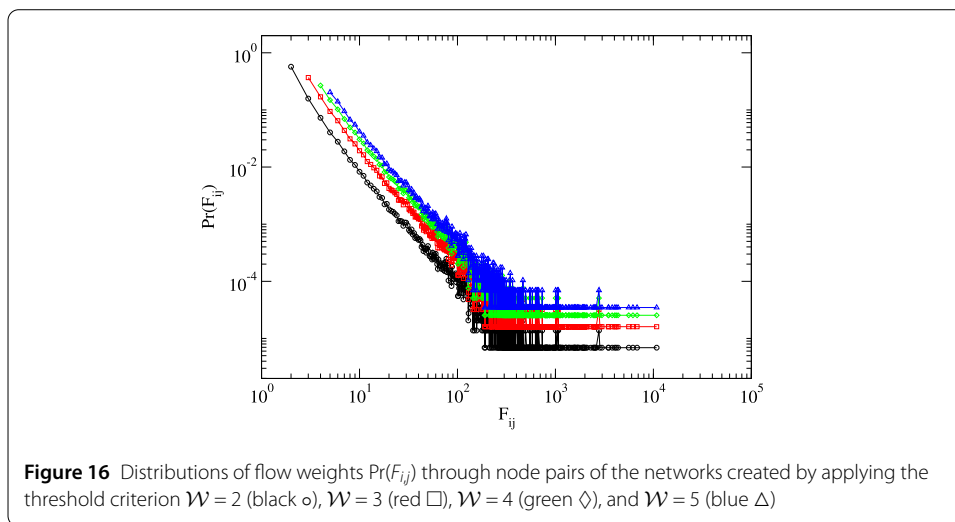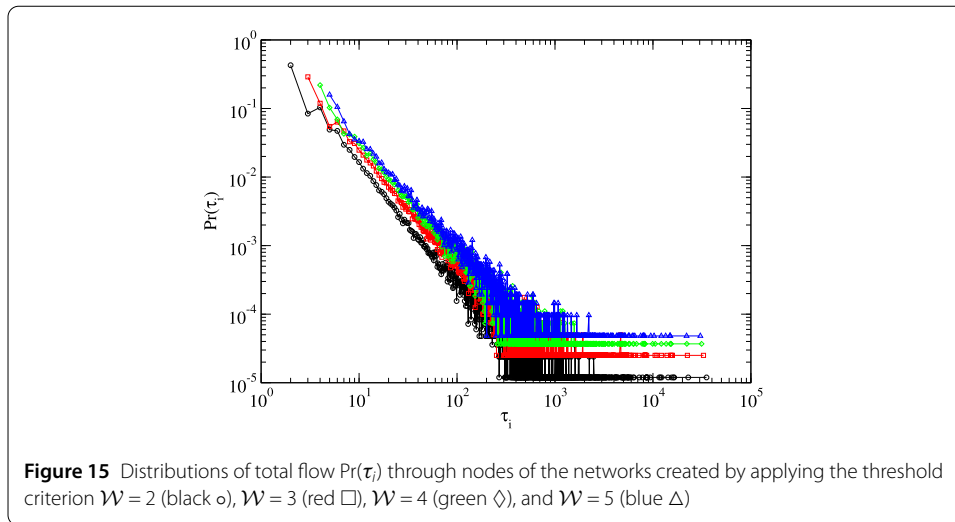
t	$\mathcal{W}$	$\bar{\mathcal{F}}_{\cdot}^{*}(\mathcal{E}_{\mathcal{W}}^{*} \cap \bar{\mathcal{E}}_{1}^{*})$	$\bar{\mathcal{F}}_{\cdot}^{*}(\mathcal{E}_{\mathcal{W}}^{*} \cap \bar{\mathcal{S}}_{1}^{*})$	$\bar{\mathcal{F}}_{\cdot}^{*}(\bar{\mathcal{E}}_{1}^{*})$	$\ln p$-value
1990	1	17,520	1762.32	44,754	−79,197.93
1995	1	17,322	1527.78	61,896	−86,403.38
2000	1	27,944	3396.18	96,933	−98,918.60
2005	1	34,484	4848.07	131,740	−101,314.57
1990	2	10,979	1375.66	28,013	−39,196.11
1995	2	10,217	1295.19	33,601	−35,239.02
2000	2	16,522	3009.81	52,207	−36,770.46
2005	2	20,864	2320.71	69,983	−83,855.28
1990	3	8298	1015.37	20,547	−32,485.19
1995	3	7942	949.05	24,333	−27,038.30
2000	3	12,921	2292.62	38,423	−31,750.46
2005	3	15,814	2229.20	53,088	−48,158.52

Figure 14 Distributions of degree $\Pr(k_i)$ of the networks created by applying the threshold criterion $\mathcal{W} = 2$ (black ○), $\mathcal{W} = 3$ (red □), $\mathcal{W} = 4$ (green ◇), and $\mathcal{W} = 5$ (blue △)
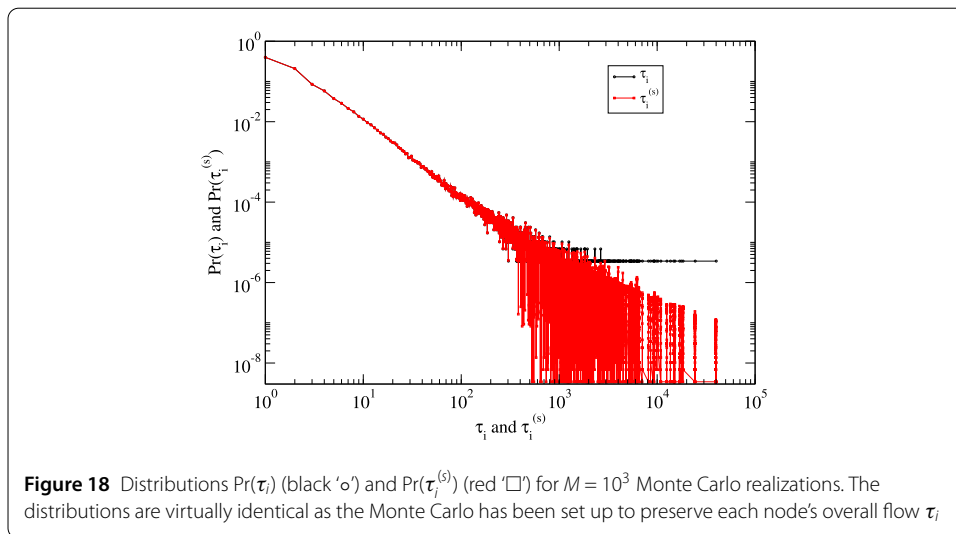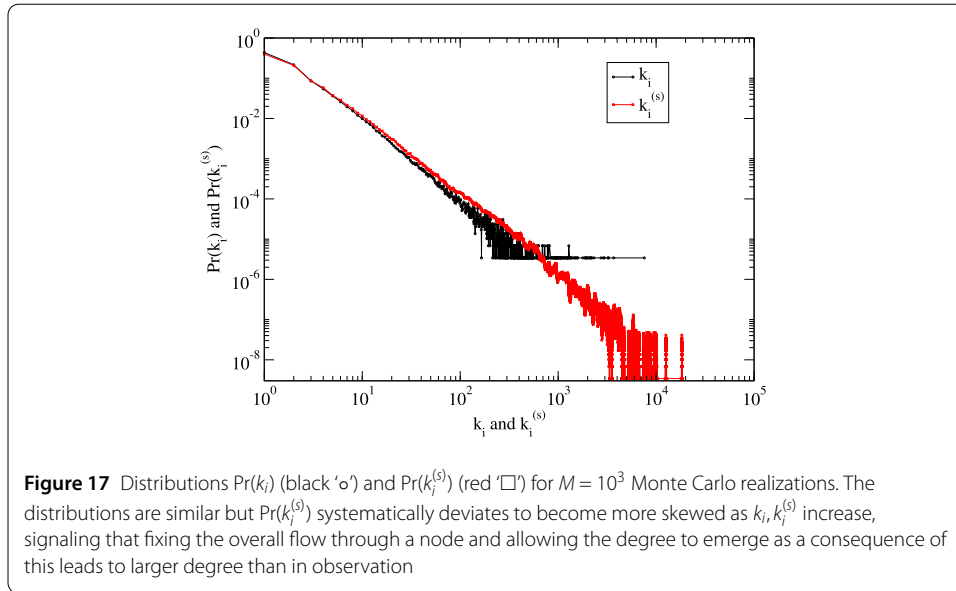
result of the τ_i transitions firm i exhibits. If this were true, the system could be explained by so called labor supply assumptions, i.e., that the sizes of firms (numbers of employees or numbers of individuals moving through the firms) determine other properties of the firms such as their degree; for the labor supply assumption to be sufficient to explain the system, transitions between firms would need to be considered random so that additional assumptions can be avoided. Thus, a configuration model reconstruction of the labor network only fixing τ_i per node would confirm (or reject) the labor supply assumption.

The counter-side to the labor supply assumption is that meaningful links between firms do exist which cannot be simply a consequence of the random jumps of individuals between firms. We have already provided evidence that indeed links are meaningful in our persistence tests above. Here we provide further evidence of the need for a labor flow network with specific links by showing that the configuration network model where only the observed τ_i are fixed does not match observation.

In Figs. 17, 18, and 19 we show, respectively, comparisons of the observed distributions of k_i, τ_i, F_{ij} with respect to their Monte Carlo simulated versions $k_i^{(s)}, \tau_i^{(s)}, F_{ij}^{(s)}$. The Monte Carlo distributions have been generated with $M = 10^3$ realizations. Distributions $\Pr(\tau_i)$

Figure 15 Distributions of total flow $\Pr(\tau_i)$ through nodes of the networks created by applying the threshold criterion $\mathcal{W} = 2$ (black ○), $\mathcal{W} = 3$ (red □), $\mathcal{W} = 4$ (green ◇), and $\mathcal{W} = 5$ (blue △)

Figure 16 Distributions of flow weights $\Pr(F_{ij})$ through node pairs of the networks created by applying the threshold criterion $\mathcal{W} = 2$ (black ○), $\mathcal{W} = 3$ (red □), $\mathcal{W} = 4$ (green ◇), and $\mathcal{W} = 5$ (blue △)

and $\Pr(\tau_i^{(s)})$ are virtually identical given the condition of the Monte Carlo to fix τ_i per node to satisfy the labor supply assumption.[g] The other distributions, however, differ between their observed and random versions. The fixed flow condition has a tendency to spread flows evenly over nodes which leads simultaneously to nodes that tend to have larger degree in simulation than in reality and links that tend to have more homogenized flows than in the real system. The effect on weights is strong enough that it is patently easy to see in the distributions. However, it is slightly harder to see the effect on k_i, but an eloquent illustration of the effect can be found in the scatter plot Fig. 20. In this plot, we see that virtually none of the degrees generated as a result of the labor flow assumption is smaller than k_i. The results systematically generated larger degree. If labor supply was properly descriptive of the real network, the cloud of generated degrees should lie evenly along the diagonal. It is of note that only approximately 0.12% of the nodes have a degree in the simulated networks that is less than the actual degree of the node; most simulated degrees are equal or larger than observed degrees. This makes clear the disparity between the networks that emerge from the labor supply assumption versus the observed network.
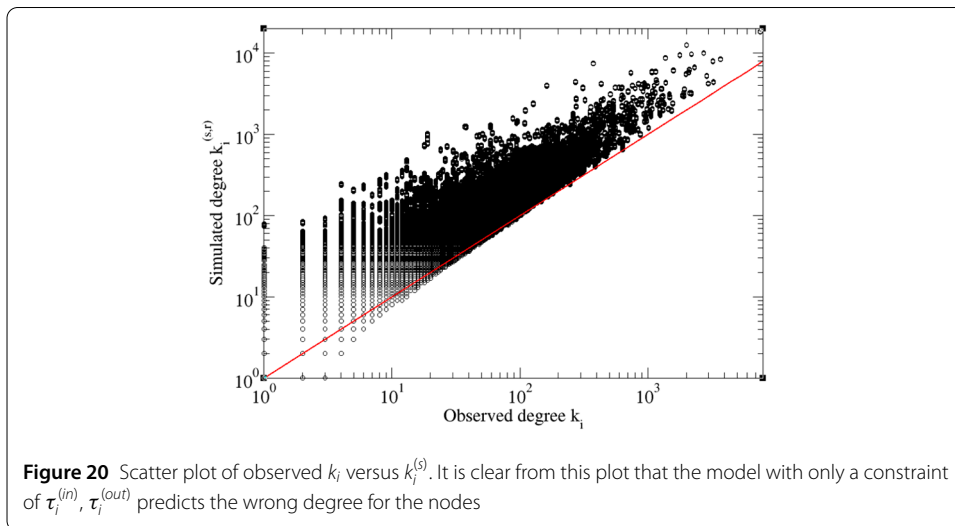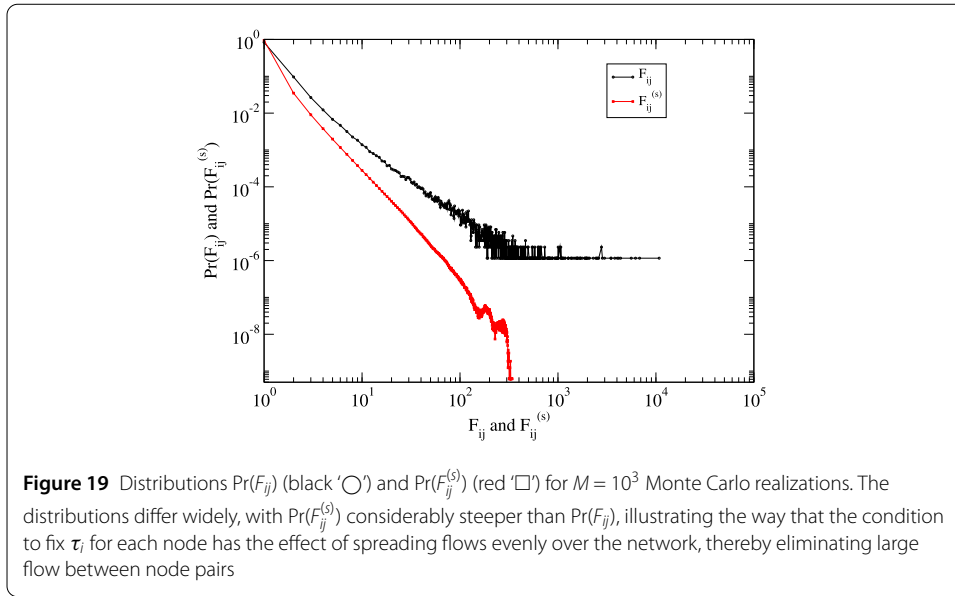
Figure 17 Distributions $\Pr(k_i)$ (black 'o') and $\Pr(k_i^{(s)})$ (red '□') for $M = 10^3$ Monte Carlo realizations. The distributions are similar but $\Pr(k_i^{(s)})$ systematically deviates to become more skewed as $k_i, k_i^{(s)}$ increase, signaling that fixing the overall flow through a node and allowing the degree to emerge as a consequence of this leads to larger degree than in observation

Figure 18 Distributions $\Pr(\tau_i)$ (black 'o') and $\Pr(\tau_i^{(s)})$ (red '□') for $M = 10^3$ Monte Carlo realizations. The distributions are virtually identical as the Monte Carlo has been set up to preserve each node's overall flow τ_i

A.5 Determining ξ_i from the data via MLE

We determine the values of ξ_i by using the fact they are the rates of success of the geometric distributions of the waiting times to be hired. Consider S_i agents that have experienced an unemployment spell from i, each with duration $t_m, m = 1, \ldots, S_i$. The log-likelihood to observe those unemployment spells is given by

$$\log\left[\prod_{\ell=1}^{S_i} \xi_i(1-\xi_i)^{t_\ell - 1}\right] = S_i\left[\log\xi_i + \left(\langle t\rangle_{S_i} - 1\right)\log(1-\xi_i)\right], \tag{38}$$

where $\langle t\rangle_{S_i} = \sum_{\ell=1}^{S_i} t_\ell / S_i$. The maximum likelihood estimator for ξ_i corresponds to

$$\xi_i = 1/\langle t\rangle_{S_i}. \tag{39}$$

Figure 19 Distributions $\Pr(F_{ij})$ (black '◯') and $\Pr(F_{ij}^{(s)})$ (red '☐') for $M = 10^3$ Monte Carlo realizations. The distributions differ widely, with $\Pr(F_{ij}^{(s)})$ considerably steeper than $\Pr(F_{ij})$, illustrating the way that the condition to fix τ_i for each node has the effect of spreading flows evenly over the network, thereby eliminating large flow between node pairs

Figure 20 Scatter plot of observed k_i versus $k_i^{(s)}$. It is clear from this plot that the model with only a constraint of $\tau_i^{(in)}$, $\tau_i^{(out)}$ predicts the wrong degree for the nodes

Thus, to determine ξ_i, we calculate the average length of time agents that last worked at i spend waiting to get their next job.

A.6 Unique solutions of the model

Equations (7) and (8) of the main text constitute a homogeneous system of linear equations for the steady state probabilities of being employed at firm i. Repeating the equations here, we have

$$\mathbf{\Lambda X} = \mathbf{0} \tag{40}$$

with

$$\mathbf{\Lambda}_{ij} = \mathbf{A}_{ji} \frac{h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}|}{\sum_{\{\gamma_j \neq \emptyset\}} \langle h \rangle_{\gamma_j} \Pr(\gamma_j)} - \delta[i,j] \tag{41}$$

and

$$\mathbf{X}_i = \lambda_i r_\infty(i), \tag{42}$$

where $\mathbf{A}$ is the adjacency matrix of the LFN. A homogeneous system such as the one above always has, at least, the trivial solution $\mathbf{X} = \mathbf{0}$. For there to be interesting, non-trivial solutions, it is necessary for the matrix $\mathbf{\Lambda}$ to be singular, i.e., to have determinant equal to zero [39].

In this section we show that $\mathbf{\Lambda}$ is indeed singular and, furthermore, that for a connected network with adjacency matrix $\mathbf{A}$, the singularity stems from matrix $\mathbf{\Lambda}$ having a reduced rank of $N - 1$, where N is the number of network nodes (matrix $\mathbf{\Lambda}$ is $N \times N$). This means that there is a 1-dimensional space of solutions for $\mathbf{X}$ which satisfies $\mathbf{\Lambda X} = \mathbf{0}$. To obtain a unique solution, one simply needs to make use of the normalization condition

$$\sum_{i=1}^{N} \big(r_\infty(i) + s_\infty(i) \big) = 1 \tag{43}$$

together with [4] of the main text,[h] which makes Eq. (43) take on the form

$$\sum_{i=1}^{N} \left[\frac{1}{\lambda_i} + \frac{1}{\sum_{\{\gamma_i \neq \emptyset\}} \langle h \rangle_{\gamma_i} \Pr(\gamma_i)} \right] \lambda_i r_\infty(i) = 1. \tag{44}$$

Before we embark on showing the proofs, we should note the intuitive reason why $\mathbf{\Lambda}$ is not of full rank, but instead has a rank reduced by one unit: since the probability flow is conserved, it is not necessary to know the probabilities at all nodes. Clearly, due to conservation, if we know the probabilities in $N - 1$ nodes, then the probability for the Nth node can be determined from Eqs. (43) or (44). Matrix $\mathbf{\Lambda}$ encodes the way the probabilities flow across the system without normalizing them.

To show first that $\mathbf{\Lambda}$ is singular, it is sufficient to show that all of its columns add to zero, which is equivalent to saying that at least one of its rows is linearly dependent on other rows. This can be seen if we first sum $\mathbf{\Lambda}_{ij}$ over i

$$\sum_{i=1}^{N} \mathbf{\Lambda}_{ij} = -1 + \sum_{i=1}^{N} \mathbf{A}_{ji} \frac{h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}|}{\sum_{\{\gamma_j \neq \emptyset\}} \langle h \rangle_{\gamma_j} \Pr(\gamma_j)}, \tag{45}$$

where -1 comes from $-\sum_i \delta[i,j]$. We can now show that the numerator and denominator of the second term are indeed equal. To see this in detail, we organize the elements of $\{\gamma_j^{(i)}\}$ by cardinality $|\gamma_j^{(i)}|$, and rewrite the numerator as

$$\sum_{i=1}^{N} \mathbf{A}_{ji} h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}| = \sum_{c=1}^{|\Gamma_j|} \frac{1}{c} \sum_i \mathbf{A}_{ji} h_i \sum_{\{|\gamma_j^{(i)}|=c\}} \Pr(\gamma_j^{(i)}), \tag{46}$$

where the last sum is over all elements of $\{\gamma_j^{(i)}\}$ with equal size c. Now, the sum over i guarantees that each neighbor of j belonging to a particular $\gamma_j^{(i)}$ is summed, along with the

corresponding h_r, where $r \in \gamma_j^{(i)}$. Therefore, the sum over i can be rewritten as

$$\sum_i \mathbf{A}_{ji} h_i \sum_{\{|\gamma_j^{(i)}|=c\}} \Pr(\gamma_j^{(i)}) = \sum_{\{|\gamma_j|=c\}} \left(\sum_{r\in\gamma_j} h_r\right) \Pr(\gamma_j) \tag{47}$$

and inserting this into the sum over c leads to

$$\sum_{c=1}^{|\Gamma_j|} \frac{1}{c} \sum_{\{|\gamma_j|=c\}} \left(\sum_{r\in\gamma_j} h_r\right) \Pr(\gamma_j) = \sum_{\{\gamma_j\neq\emptyset\}} \frac{\sum_{r\in\gamma_j} h_r}{|\gamma_j|} \Pr(\gamma_j) = \sum_{\{\gamma_j\neq\emptyset\}} \langle h\rangle_{\gamma_j} \Pr(\gamma_j) \tag{48}$$

Therefore,

$$\sum_{i=1}^{N} \mathbf{A}_{ij} h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}| = \sum_{\{\gamma_j\neq\emptyset\}} \langle h\rangle_{\gamma_j} \Pr(\gamma_j), \tag{49}$$

which means that for all j, (45) is identically zero.

In order to determine the rank of matrix $\mathbf{\Lambda}$, we first highlight the following relations to the flow probabilities of the random walker. First, note that

$$h_i s(j,t) \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}| = \eta_{ji}, \tag{50}$$

is the probability current for the transition $s(j,t) \to r(i,t+1)$ from unemployment at node j to employment at node i. Also,

$$s(j,t) \sum_{\gamma_j\neq\emptyset} \Pr(\gamma_j)\langle h\rangle_{\gamma_j} = \eta_{j.}, \tag{51}$$

is the entire probability flow out of j, which is to say $s(j,t) \to r(\Gamma_j, t+1)$ (i.e., the probability of being employed at any of the neighbors of j). Due to conservation of probability, it is clear that

$$\eta_{j.} = \sum_{i\in\Gamma_j} \eta_{ji}, \tag{52}$$

which we use below.

The proof that $\mathbf{\Lambda}$ has incomplete rank can also be written by making use of Eq. (52). Hence

$$\sum_{i=1}^{N} \mathbf{A}_{ji} h_i \sum_{\{\gamma_j^{(i)}\}} \Pr(\gamma_j^{(i)})/|\gamma_j^{(i)}| = \frac{\sum_{i\in\Gamma_j} \eta_{ji}}{s(j,t)} = \frac{\eta_{j.}}{s(j,t)} = \sum_{\gamma_j\neq\emptyset} \Pr(\gamma_j)\langle h\rangle_{\gamma_j}. \tag{53}$$

In addition to this simplification, one can also make use of this reformulation to test for the rank of a reduced matrix $\mathbf{\Lambda}'$ which is equivalent to $\mathbf{\Lambda}$ but missing a row and a column. In this case, we choose to eliminate the Nth row/column, which is a fully general choice given the arbitrary nature of the labeling of nodes.

We want to show that, for a connected graph, the rows of $\mathbf{\Lambda}'$ form a linearly independent set of vectors. Thus, defining

$$\vec{\Lambda}_i = (\Lambda_{i,1},\ldots,\Lambda_{i,N-1}), \quad [i = 1,\ldots,N-1], \tag{54}$$

where each Λ_{ij}, when rewritten by using Eqs. (50) and (51), is

$$\Lambda_{ij} = \mathbf{A}_{ji}\frac{\eta_{ji}}{\eta_{j.}} - \delta[i,j], \tag{55}$$

we can form an equation on the coefficients $\{\mu_i\}$

$$\vec{0} = \sum_{i=1}^{N-1} \mu_i \vec{\Lambda}_i \tag{56}$$

and show that it can only be satisfied if all μ_i are zero, which would prove that the set of vectors $\vec{\Lambda}_1,\ldots,\vec{\Lambda}_{N-1}$ are linearly independent. Expanding the right hand side of this last expression, we obtain

$$\sum_{i=1}^{N-1} \mu_i \vec{\Lambda}_i = \left(\sum_{i=1}^{N-1} \mu_i \Lambda_{i1}, \sum_{i=1}^{N-1} \mu_i \Lambda_{i2}, \ldots, \sum_{i=1}^{N-1} \mu_i \Lambda_{i,N-1}\right)$$
$$= \left(\frac{\sum_{i=1}^{N-1} \mu_i \mathbf{A}_{1i}\eta_{1i}}{\eta_{1.}} - \mu_1, \ldots, \frac{\sum_{i=1}^{N-1} \mu_i \mathbf{A}_{N-1,i}\eta_{N-1,i}}{\eta_{N-1.}} - \mu_{N-1}\right). \tag{57}$$

Equating this to $\vec{0}$, we obtain the set of equations

$$\frac{\sum_{i\in\Gamma'_j} \mu_i\eta_{ji}}{\eta_{j.}} - \mu_j = 0 \quad [\forall j = 1,\ldots,N-1], \tag{58}$$

where Γ'_j represents the graph neighbors of j excluding node N. With the aid of Eq. (52), this set of equations can be rewritten as

$$\sum_{i\in\Gamma'_j} \mu_i\eta_{ji} - \mu_j \sum_{i\in\Gamma_j} \eta_{ji} = 0 \quad [\forall j = 1,\ldots,N-1], \tag{59}$$

where one should note that for $\eta_{j.}$, the set of neighbors used includes N. Explicitly separating the (possible) flow between j and N, this becomes

$$\sum_{i\in\Gamma'_j} \mu_i\eta_{ji} - \mu_j \sum_{i\in\Gamma'_j} \eta_{ji} = \mu_j\mathbf{A}_{j,N}\eta_{j,N}$$
$$\Rightarrow \sum_{i\in\Gamma'_j} (\mu_i - \mu_j)\eta_{ji} = \mu_j\mathbf{A}_{j,N}\eta_{j,N} \quad [\forall j = 1,\ldots,N-1]. \tag{60}$$

The last expression highlights the relationships that need to hold among the otherwise independent link flows to satisfy Eq. (56). From a physical standpoint, these relations tell us that the vectors in Eq. (54) are indeed linearly independent in contrast to the case when

all the nodes are present (i.e., the full matrix $\mathbf{\Lambda}$ as opposed to the reduced matrix $\mathbf{\Lambda}'$). In matrix $\mathbf{\Lambda}$, there is redundant information because knowing all the flows for $N - 1$ nodes provides the information needed to determine the flows to the remaining node. But for the row vectors of $\mathbf{\Lambda}'$, that is Eq. (54), the absence of the variables accounting for the flows going into and out of one of the nodes (in our case N) breaks the linear dependence. We see this in more detail next.

In order to find the $\{\mu_i\}$ that solve Eq. (56), or equivalently Eq. (60), we proceed as follows. Out of the $N - 1$ members of Eq. (60), consider the ones where j is not a neighbor of N. This means that $\mathbf{A}_{j,N} = 0$, making the right hand sides equal to 0. It also means that, for this case, $\Gamma'_j = \Gamma_j$ because there are no connections to N. Then, the only admissible solutions require that $\mu_i = \mu_j$ (which we label μ for simplicity) for all $i \in \Gamma_j$, since the flows $\eta_{ji} \neq 0$. In other words

$$\sum_{i \in \Gamma_j} (\mu_i - \mu_j)\eta_{ji} = \sum_{i \in \Gamma_j}(\mu - \mu)\eta_{ji} = 0 \quad [j \notin \Gamma_N]. \tag{61}$$

Furthermore, we note that the μ chosen for one j "propagates" to other nodes since the network is connected (has a single cluster). To understand this, let us consider two situations for i: 1) i is itself not a neighbor of N, in which case it also needs to satisfy Eq. (61), but because this i is connected to j, also satisfying Eq. (61), then they must share the same μ value for consistency, or 2) i is a neighbor of N (which we consider next). Before we tackle 2), note that 1) implies all the neighbors of N (Γ_N) also have $\mu_j = \mu$ because the nodes that are at distance 2 from N all propagate μ to the nodes in Γ_N.

When a node j is a neighbor of N, $\mathbf{A}_{j,N} = 1$, and it is clearer to write Eq. (60) in the form (59), and use Eq. (52) to obtain

$$\sum_{i \in \Gamma'_j} \mu_i \eta_{ji} - \mu_j \eta_{j.} = 0 \quad [j \in \Gamma_N]. \tag{62}$$

But because all the μ_i above have been shown to be equal to μ, we have

$$\mu \sum_{i \in \Gamma'_j} \eta_{ji} - \mu\eta_{j.} = \mu(\eta_{j.} - \eta_{j,N}) - \mu\eta_{j.} = -\mu\eta_{j,N} = 0 \quad [j \in \Gamma_N], \tag{63}$$

which cannot be generally satisfied unless $\mu = 0$, showing that indeed the only solution for Eq. (56) is $\mu_i = 0, \forall i = 1, \ldots, N$.

If all nodes are neighbors of N, Eq. (60) for all nodes is given by

$$\sum_{j \in \Gamma'_j} (\mu_i - \mu_j)\eta_{ji} = \mu_j\eta_{j,N} \tag{64}$$

but because the η_{ji} are independent, the only way to satisfy the equation is if $\mu_i = 0$ for all i.

A.7 Robustness of results to changes in the $\mathcal{W}$ threshold

In the main text and in Sect. A.3, we discuss how we construct the network encoded in $\mathbf{A}$. We indicate that the edges are considered significant if multiple transitions take place between the two firms connected by the edge, and we set the minimum number of transitions

to $\mathcal{W} = 2$. This choice ensures maximum amount of data to increase statistical significance of the results.

However, for this threshold to be acceptable, we have to check whether our results are robust to increasing $\mathcal{W}$. This is the purpose of this section and, as we indicated in the main text, the results below support our choice to use $\mathcal{W} = 2$ for our main analysis. To undertake the robustness test, we construct plots of the quantity $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ which are equivalent to Fig. 2 of the main text, but here we use the minimum number of job-to-job transitions to be $\mathcal{W} = 3, 4, 5$ (Fig. 21). If the key characteristics of the surface $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ remain as $\mathcal{W}$ increases, it means that the results for $\mathcal{W} = 2$ are not an artifact of such a choice. The key characteristic we care about is the typical functional relation between L_i^* and k_i/λ_i. Just as for the plot discussed in the main text, we can see that the level sets of maximum probability are parallel to the linear "best fit" planes. Note that in all the plots, for values of $L_i \sim 10^2$ and above together with $k_i/\lambda_i \sim 10^2$ and above, the level sets of highest probability run parallel to the linear planes of "best fit". Furthermore, we can observe that as $\mathcal{W}$ increases, the range of the level sets that deviates from the "best fit" planes is located at the small values of L_i^* and k_i/λ_i, which should not be surprising since an increase in $\mathcal{W}$ means that firms with very few numbers of employees cannot be well represented.

Therefore, the overall conclusion is that the match between the homogeneous model, represented by Eqs. (15) and (16) with $r_\infty(i) = r_\infty^{(v)}(i)$ and $s_\infty(i) = s_\infty^{(v)}(i)$ from Eqs. (11) and (12), respectively, are supported by our analysis of the Finnish data even as the criterion for including a link in **A** is made more stringent.

A.8 Further evidence of steady-state behavior

While the balanced-flows plot presented in the main text provide evidence of steady-state behavior in the data, the reader might be concerned about other types of fluctuations that are not addressed by such test. For example, it might be the case that, even with balanced flows and a constant population, the firm size cumulative distribution might change its shape through time. Figure 22 shows the firm size cumulative distribution obtained for yearly measurements from the Finnish data. It is clear that the distribution is robust across time, and that the only changes are upward shifts due to the population growth through several years. This way of showing evidence of steady-state behavior is more standard in the study of firm dynamics.

In addition to the previous test, we also check that the effects that may be proportional to firm size do not hide systematic behavior that we cannot see in the simple study of raw flows. In Fig. 23, we plot the histogram of the net fractional flow for Finland. Here too, we see a very large concentration of the histogram around a zero fractional flow. Note the vertical logarithmic scale.

As a final test, we look at the changes in the network structure from year to year using the time dimension of our data. Consider $V(t)$ the set of network links that are accepted to the overall labor flow network when the data is analyzed starting from the year t and going to the end of the data in year t_f (the data spans from $t_o = 1988$ to $t_f = 2007$). In other words, the links in $V(t)$ are between node pairs that had flow at or above the threshold $\mathcal{W} = 2$ between years t and t_f. Now, we calculate $V(t)$ for all possible t (i.e. systematically varying the starting year) which creates the series $V(t_o), V(t_o + 1), \ldots, V(t_f - 1)$. From these
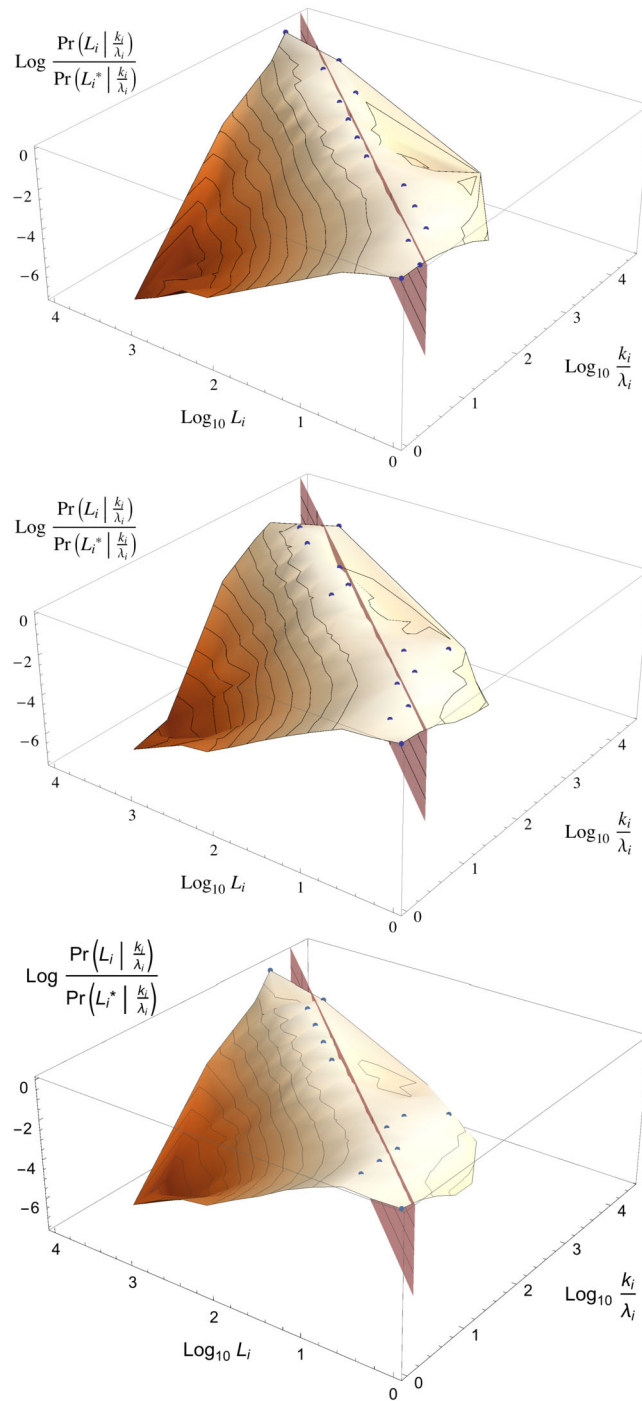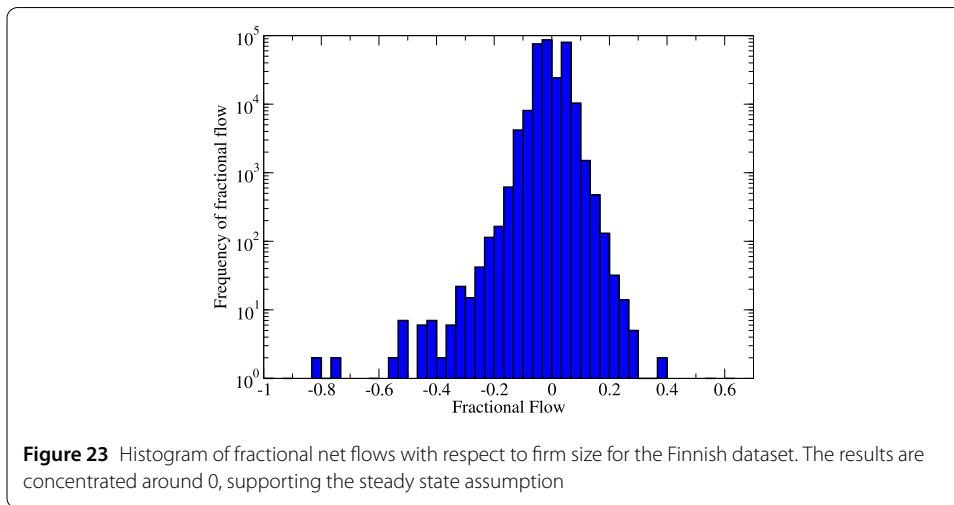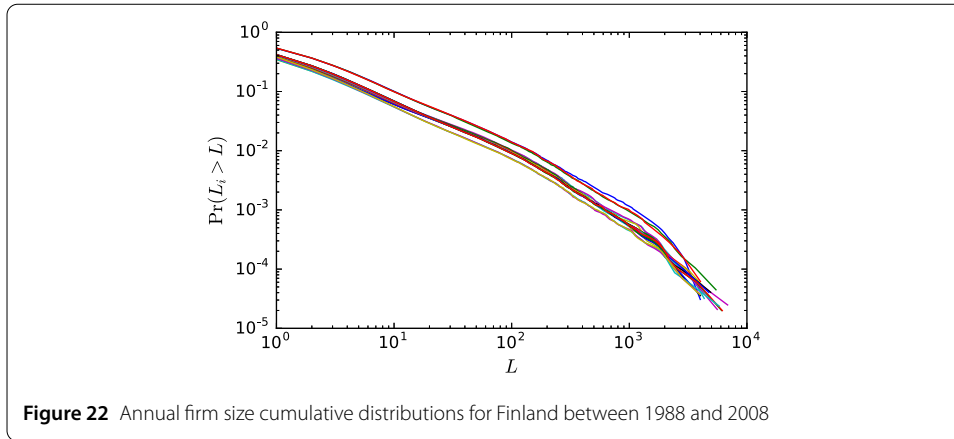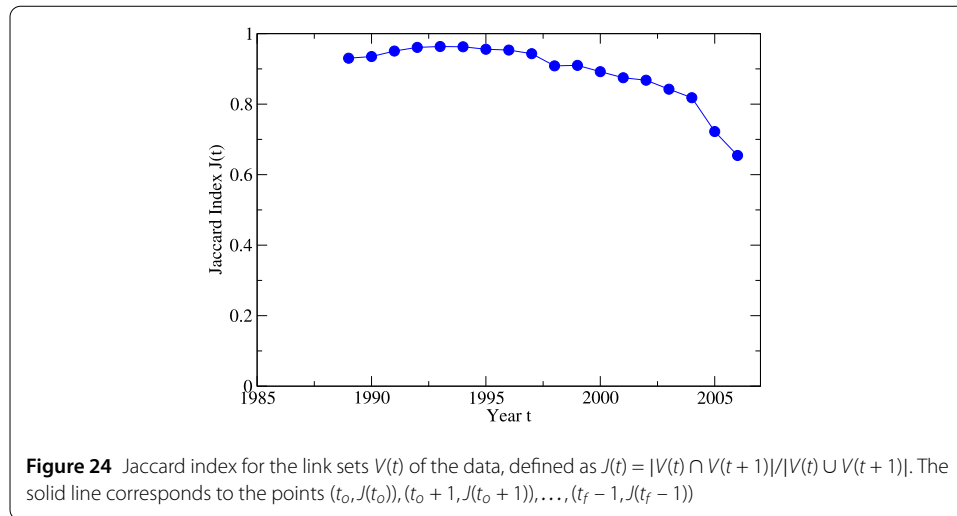
Figure 21 Behavior of $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_j^*|k_i/\lambda_i)]$ (surface $\mathcal{S}$) with respect to $\log_{10} L_i$ and $\log_{10} k_i/\lambda_i$ for Finland when **A** is created with links where the traffic has a minimum of 3,4,5 transitions (top, middle, bottom plots). The data is logarithmically binned as follows: L_i belongs to bin b (a non-negative integer) if $L_{\mathbf{min}}\zeta^b < L_i \leq L_{\mathbf{min}}\zeta^{b+1}$ with $\zeta > 1$ (for this plot $\zeta = 2$) and $L_{\mathbf{min}} = \min[\{L_i\}]$ (smallest firm size in the data); k_i/λ_i is binned in the same way with ζ and $(k/\lambda)_{\mathbf{min}} = \min[\{(k_i/\lambda_i)\}]$. Blue points represent the local maximum of $\mathcal{S}$ at each bin. The vertical plane $\mathcal{P}$ (linear "best fit") is parametrized as $(k_i/\lambda_i, C_L k_i/\lambda_i, z)$ where z is a free parameter. C_L is chosen to minimize $\sum_b (L_b^* - C_L(k/\lambda)_b)^2$ with the first three bins excluded because the smallest firm size is 1. The large range within which the intersection of $\mathcal{P}$ and $\mathcal{S}$ runs parallel to the maxima of $\mathcal{S}$ strongly supports Eq. (18) of the main text

Figure 22 Annual firm size cumulative distributions for Finland between 1988 and 2008

Figure 23 Histogram of fractional net flows with respect to firm size for the Finnish dataset. The results are concentrated around 0, supporting the steady state assumption

results, we define the Jaccard index $J(t)$ of the links as the quotient

$$J(t) = \frac{|V(t) \cap V(t+1)|}{|V(t) \cup V(t+1)|},$$
(65)

where the numerator counts how many links that are accepted into $V(t)$ from analyzing the data from year t to t_f are also accepted into $V(t+1)$ using years $t+1$ to t_f of the data, and the denominator counts the number of unique links that belong to either $V(t)$ or $V(t+1)$ including links belonging to both. In other words, $J(t)$ measures how similar $V(t)$ and $V(t+1)$ are to one another; at its limit values, $J(t)$ is equal to 1 when $V(t) = V(t+1)$ and 0 when $V(t) \cap V(t+1) = \emptyset$, i.e., $V(t)$ and $V(t+1)$ do not have any common links. Figure 24 shows the results of this analysis. For the vast majority of years, the Jaccard index is close to 1 and decays slightly towards the later years due to the short time range available to test links against the acceptance criterion. The meaning of these results is that despite the dynamics of the network, the labor flow network determined from persistent links changes by 10% or less in any typical year. Thus our method should be stable to yearly changes in the network over time to a level that generally exceeds 90%.

Figure 24 Jaccard index for the link sets $V(t)$ of the data, defined as $J(t) = |V(t) \cap V(t+1)|/|V(t) \cup V(t+1)|$. The solid line corresponds to the points $(t_o, J(t_o)), (t_o + 1, J(t_o + 1)), \ldots, (t_f - 1, J(t_f - 1))$
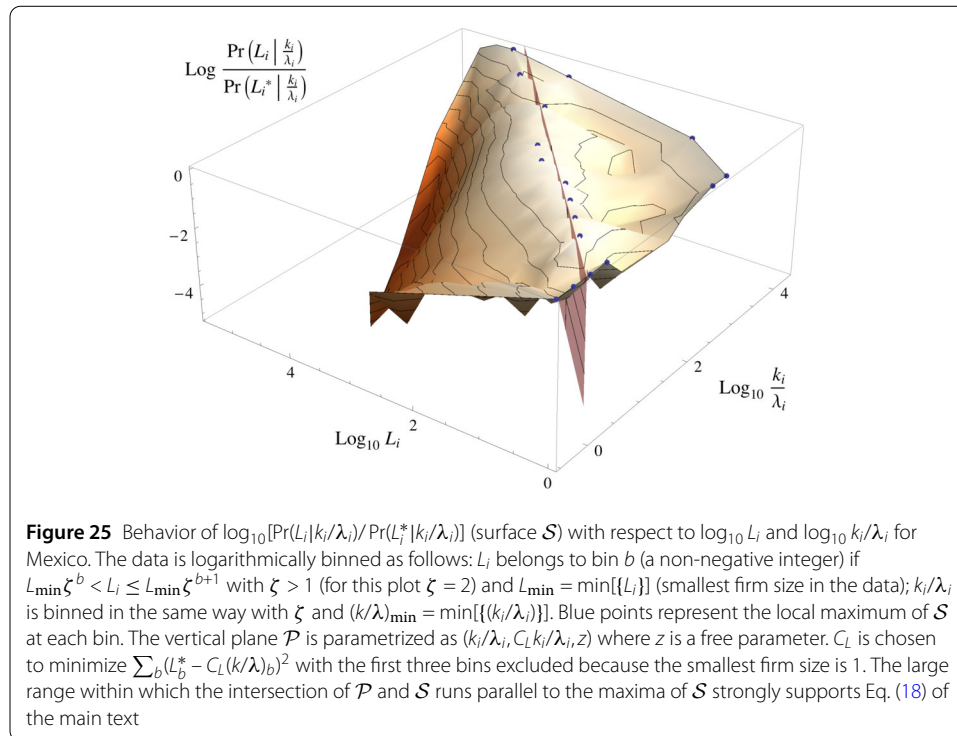
A.9 Firm sizes for Mexican data

In the main text, we present the analysis of firm sizes for Finland and the firm-specific unemployment for Mexico but, because there is no data available to deal with Finnish firm-specific unemployment, we do not present results for this case. However, there is data for firm sizes for Mexico which, in the interest of space, we do not show in the main text. Here, we present this analysis as a supplement to our main analysis. Because of the method of collection of the data for Mexico, it is not possible to have results as clean as those for Finland since the sampling has not been done uniformly on firms but rather on individuals. However, the results for the Mexican data are not inconsistent with the Finnish results.

As in the main text, we use the data to draw a comparison with the homogeneous model that assumes $v_i = v$. This model predicts that $L_i \sim k_i/\lambda_i$, as seen from Eq. (18) of the main article. Therefore, in Fig. 25, we present the 3-dimensional plot of $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ as a function of $\log_{10} L_i$ and $\log_{10}(k_i/\lambda_i)$. Once again, $\Pr(L_i^*|k_i/\lambda_i)$ corresponds to the probability associated with the conditional mode L_i^*. The normal plane is parametrized as indicated in the caption. The relation between L_i and k_i/λ_i is again supported by this plot, although more fluctuations appear towards firms of larger size and/or of large degree k_i. These firms are less likely to be properly sampled because, even though they have more employees, there are fewer of them (the same complication emerges for the plots of firm-specific unemployment, by the way).

A.10 RANSAC

The RANSAC algorithm [37] is a popular method to estimate parametric models when data has large amounts of outliers, which is our case. The algorithm performs an OLS estimation by randomly sampling the minimum number of data points needed to fit the model. Then, it identifies those data points from the entire dataset that fall within a given distance from the estimated model. These two steps are repeated until a model with a maximum number of inlier data points is found. Since RANSAC is non-deterministic, it yields slightly different estimations every time it is performed. For this reason, we performed 1 million RANSAC estimations and analyzed the distribution of their estimates. We employed the default parameters of 2 minimum observations, the median absolute deviation

Figure 25 Behavior of $\log_{10}[\Pr(L_i|k_i/\lambda_i)/\Pr(L_i^*|k_i/\lambda_i)]$ (surface $\mathcal{S}$) with respect to $\log_{10} L_i$ and $\log_{10} k_i/\lambda_i$ for Mexico. The data is logarithmically binned as follows: L_i belongs to bin b (a non-negative integer) if $L_{\mathbf{min}}\zeta^b < L_i \le L_{\mathbf{min}}\zeta^{b+1}$ with $\zeta > 1$ (for this plot $\zeta = 2$) and $L_{\mathbf{min}} = \min[\{L_i\}]$ (smallest firm size in the data); k_i/λ_i is binned in the same way with ζ and $(k/\lambda)_{\mathbf{min}} = \min[\{(k_i/\lambda_i)\}]$. Blue points represent the local maximum of $\mathcal{S}$ at each bin. The vertical plane $\mathcal{P}$ is parametrized as $(k_i/\lambda_i, C_L k_i/\lambda_i, z)$ where z is a free parameter. C_L is chosen to minimize $\sum_b (L_b^* - C_L(k/\lambda)_b)^2$ with the first three bins excluded because the smallest firm size is 1. The large range within which the intersection of $\mathcal{P}$ and $\mathcal{S}$ runs parallel to the maxima of $\mathcal{S}$ strongly supports Eq. (18) of the main text

as a threshold to distinguish between inliers and outliers, and the maximum number of trials of 100 which are standard values in the literature.

To be specific, each RANSAC run consists of two steps. First, a data subset is sampled and the model is estimated. The sample size has to be of the minimum necessary to estimate the model. Second, the data that was not sampled is compared against the predictions of the estimated model. Given an error threshold, those data points that are not predicted by the model are considered outliers. The points that are correctly predicted form the consensus set. If the consensus set contains too few data points, the estimated model is dropped and the algorithm goes back to step one. Otherwise, the model is re-estimated with the consensus set and the error is computed. These steps are repeated until the error term reaches a threshold. Parameter calibration used default values from the implementation in the SciKit Learn package.

Abbreviations
FLEED, Finnish Longitudinal Employer–Employee Data; FSD, Firm Size Distribution; LFN, Labor Flow Network; OLS, Ordinary Least Squares; RANSAC, Random Sample Consensus.

Availability of data and materials
The data used in this article is directly available from the sources, Tilastokeskus (Statistics Finland), and the Instituto Mexicano del Seguro Social (Mexican Social Security Institute).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
EL and OAG designed research; EL performed calculations; EL and OAG analyzed data; EL, OAG, and RAA wrote manuscript. All authors read and approved the final manuscript.

Author details
[1] Department of Computational and Data Sciences, George Mason University, 4400 University Drive, MS 6A2, Fairfax, VA, USA. [2] Green Templeton College, University of Oxford, 43 Woodstock Rd, OX2 6HG, Oxford, UK. [3] The Alan Turing Institute, 96 Euston Rd, Kings Cross, NW1 2DB, London, UK. [4] Department of Economics, University College London, Drayton House, 30 Gordon St, Kings Cross, WC1H 0AX, London, UK.

Endnotes
[a] In fact, beyond the temporal tests presented in this Sec., we conducted non-temporal tests not reported here. However, these tests are inappropriate for this context because they have a tendency to suggest that flows between small firms are significantly non-random and those for large firms are, instead, more likely to be random. Crucially, every time that one finds a pair of nodes with a single flow event, such tests suggest strong significance, even though this flow may never be seen again, undermining the entire meaning of a network of steady flows.

[b] The increase of $\wp$ as a function of $\mathcal{W}$ may seem surprising at first. However, we must remember that $\wp$ depends on $\bar{\mathcal{E}}_1^*$ and $\mathcal{N}_{\mathcal{W}}^*$ (and on $\mathcal{E}_{\mathcal{W}}^*$, through $\mathcal{N}_{\mathcal{W}}^*$), and thus as $\mathcal{W}$ increases, the density of the resulting set of flows among the pairs of nodes in $\mathcal{N}_{\mathcal{W}}^*$ increases because the nodes selected tend to be those with more overall flows anyway. Note that the increases of $\wp$ with respect to $\mathcal{W}$ appear more pronounced than those of $\wp_{\mathcal{W}}$, but this is because $\wp_{\mathcal{W}}$ is already large and starts to have less room for increase as it begins to approach 1 (a saturation effect).

[c] The parameters that control the stochastic model, defined in the main text, can be systematically chosen on the basis of well structured economic models that optimize certain economic features (see e.g. [15])

[d] For a connected graph, the solution is unique, but if the graph is separated into disconnected components, one can set up calculations identical to the one here for each component and obtain a unique result for the set of all components.

[e] In this manuscript, prediction does not refer to temporal forecasting. Rather, it specifically means the result of solving our model and using it to calculate quantities that can be contrasted with their empirically measured equivalents.

[f] Even more surprisingly, the estimate for α to be close to 1 is usually difficult to obtain when RANSAC is performed without restriction. The usual procedure is to *assume* α has a given value and estimate β only, but in our case, this is not necessary.

[g] Although most of the distribution of $\tau_i^{(s)}$ is identical to that of τ_i, there are differences at the right tail. This is because the algorithm for rewiring the flows sometimes produces networks with a small number of unmatched flows (in a sample of $M = 10^3$ rewired networks, an average of 919.5 ± 30.3 flows were missed out of $\approx 1.8 \times 10^6$ for each random network). This happens when the algorithm tries to rewire a node with itself. If this happens, we simply ignore this attempt and move onto the next match of in- and out-flows. The probability of this happening is very small, hence leading to only a fraction of about 5×10^{-4} missed flows. Large nodes are mores susceptible to this. Around $k_x = 750$, each node i with $k_i \geq k_x$ is likely to have at least one missed flow per realization. There are only 97 such nodes in the network of 292,614 nodes. The average number of flows missed among those nodes is less than 10 per node.

[h] Using [4] from the main text allows us to write $s_\infty(i)$ in terms of $r_\infty(i)$, and hence have the normalization condition written purely in terms of **X**.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
 1. Parsons T (1937) The structure of social action, volume I. Marshall, Pareto. McGraw-Hill Book Company, New York
 2. Hedström P, Bearman P (eds) (2009) The Oxford handbook of analytical sociology Oxford University Press, Oxford
 3. Kydland FE, Prescott EC (1982) Time to build and aggregate fluctuations. Econometrica 50(6):1345–1370
 4. Epstein JM, Axtell RL (1996) Growing artificial societies. Brookings Institution Press & MIT Press, Washington Dc
 5. Simon H (1996) The sciences of the artificial. MIT Press, Cambridge
 6. Dosi G, Fagiolo G, Roventini A (2010) Schumpeter meeting Keynes: a policy-friendly model of endogenous growth and business cycles. J Econ Dyn Control 34(9):1748–1767
 7. Guerrero OA, López E (2017) Understanding unemployment in the era of big data: policy informed by data-driven theory. Policy Internet 9:28–54
 8. Petrongolo B, Pissarides C (2001) Looking into the black box: a survey of the matching function. J Econ Lit 39:390–431
 9. Mortensen D, Pissarides C (1994) Job creation and job destruction in the theory of unemployment. Rev Econ Stud 61:397–415
10. Gabaix X (2011) The granular origins of aggregate fluctuations. Econometrica 79(3):733–772
11. Garicano L, Lelarge C, Van Reenen J (2016) Firm size distortions and the productivity distribution: evidence from France. Am Econ Rev 106(11):3439–3479
12. Barinchon R, Figura A (2015) Labor market heterogeneity and the aggregate matching function. Am Econ J Macroecon 7(4):222–249
13. Maliranta M, Nikulainen T (2008) Labour force paths as industry linkages: a perspective on clusters and industry life cycles. ETLA Discussion Papers 1168

14. Guerrero OA, Axtell RL (2013) Employment growth through labor flow networks. PLoS ONE. https://doi.org/10.1371/journal.pone.0060808
15. Axtell RL, Guerrero OA, López E (2019) Frictional unemployment on labor flow networks. J Econ Behav Control 160:184–201
16. Park J, Wood IB, Jing E, Nematzadeh A, Ghosh S, Conover MD, Ahn Y-Y (2019) Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters in the world economy. Nat Commun 10:3449
17. Atalay E, Hortaçsu A, Roberts J, Syverson C (2011) Network structure of production. Proc Natl Acad Sci USA 108(13):5199–5202
18. Collet F, Hedströmb P (2013) Old friends and new acquaintances: tie formation mechanisms in an interorganizational network generated by employee mobility. Soc Netw 35:288–299
19. Guerrero OA, López E (2015) Labor flows and the aggregate matching function: a network-based test using employer-employee matched records. Econ Lett 136:9–12
20. Simon HA, Bonini CP (1958) The size distribution of business firms. Am Econ Rev 48:607–617
21. Stanley MHR, Amaral LAN, Buldyrev SV, Havlin S, Leschhorn H, Maass P, Salinger MA, Stanley HE (1996) Scaling behavior in the growth of companies. Nature 379:804–806
22. Axtell RL (2001) Zipf distribution of U.S. firm sizes. Science 293:1818–1820
23. Aoyama H, Fujiwara Y, Ikeda Y, Iyetomi H, Souma W (2010) Econophysics and companies. Cambridge University Press, Cambridge
24. Agarwal R, Ohyama A (2012) Industry or academia, basic or applied? Career choices and earnings trajectories of scientists. Manag Sci 59(4):950–970
25. Campbell BA, Ganco M, Franco AM, Agarwal R (2011) Who leaves, where to, and why worry? Employee mobility, entrepreneurship and effects on source firm performance. Strateg Manag J 33(1):65–87
26. Rogerson R, Shimer R, Wright R (2005) Search-theoretic models of the labor market: a survey. J Econ Lit 43:959–988
27. Coad A (2009) The growth of firms. A survey of theories and empirical evidence. Edward Elgar, Cheltenham Glos
28. Stendl J (1965) Random processes and the growth of firms. A study of the Pareto law. Hafner, New York
29. Gibrat R (1931) Les inégalités economiques. Sirey, Paris
30. Gabaix X (1999) Zipf's law for cities: an explanation. Q J Econ 114:739–767
31. Luttmer EGJ (2012) Technology diffusion and growth. J Econ Theory 147:602–622
32. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
33. Calvo-Armengol A, Jackson MO (2004) The effects of social networks on employment and inequality. Am Econ Rev 94(3):426–454
34. Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton
35. Gianelle C (2011) Exploring the complex structure of labour mobility networks: evidence from Veneto microdata. WP2011 13
36. Garlaschelli D, Ahnert SE, Fink TMA, Caldarelli G (2013) Optimal scales in weighted networks social informatics. Lect Notes Comput Sci 8238: 346–359
37. Fischler M, Bolles R (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24:381–395
38. Serrano MA, Boguñá M, Pastor-Satorras R (2006) Correlations in weighted networks. Phys Rev E 74:055101
39. Hoffman K, Kunze R (1971) Linear algebra, 2nd edn. Prentice-Hall, Upper Saddle River