Defining Labor Markets from Worker Flows

Ryan Boone and Tomas Guanziroli

April 2022

Abstract:

We use worker flows to define better approximations to labor markets. Labor market definitions are important to predict the effects of mergers, labor market policies (like minimum wages), migration, trade shocks, and other events on labor market outcomes. Researchers have typically relied on ad-hoc approximations such as geographic boundaries, occupational codes, or industry codes at differing levels of granularity. For any given job, it is not clear which definition is most appropriate. For example, the relevant market for airline pilots might be a detailed occupation-industry cell at the national level. On the other hand, the relevant market for pharmacy cashiers might consist of many occupations and industries but only within a narrow geographic area. The paper follows two steps. First, we use an approach from the network literature to identify labor markets. We use worker flows across occupations, industries, and geographies to define N markets. The defined markets maximize a measure of the density of links within markets. Second, we describe the resulting markets. For example, we show that labor markets often cross one-digit industry and occupation codes.

1. Introduction

Labor market definitions are important to evaluate and predict the effects of government policies, migration, trade shocks, mergers and acquisitions and other events on workers' outcomes. Researchers have typically relied on ad-hoc approximations for labor markets such as geographic boundaries, occupational codes, or industry codes at differing levels of granularity. However, markets may vary in size and may cross these boundaries. For any given job, it is not clear which definition is most appropriate.

In this paper, we use job transitions across county by industry by occupation cells to draw labor market boundaries in Brazil. We use a method from the network literature—the constant Potts model (CPM)—that aggregates cells (hereafter, nodes) into communities such that most job transitions occur within communities. We refer to communities as data-driven labor markets. We acknowledge that the jobs available to two distinct workers will not have a perfect overlap, such that one should interpret the data-driven labor markets as approximations to labor markets and an improvement from ad-hoc definitions.

We use the Brazilian employer-employee matched dataset (RAIS) between 2007 and 2013 to compile job transitions across nodes. The Brazilian data has many advantages. First, RAIS is a large dataset with detailed geographic, industry, and occupation categories. This allows us to define nodes at the finest level and still have a large number of job transitions across nodes. Second, Brazil is a large country—comparable in size to the US—allowing us to study geographic dispersion. Brazil is also a relatively closed economy, where most workers never worked in another country. It is reasonable to assume that labor markets will not cross international borders, and hence, will be captured with the RAIS data. A set back from RAIS is that it does not include workers in informal jobs.

The method is successful in its purpose of creating communities that are densely linked. By using 40 million job transitions across 246 thousand nodes, the algorithm finds 54,855 data-driven labor markets. We use the Leiden algorithm (Traag et al, 2018) to implement the CPM method. This algorithm corrects a flaw in another widely used algorithm in the networks literature that resulted in communities that were disconnected. In addition, the CPM method allows us to identify small communities. The data driven markets have 75% of job transitions occurring within markets and self-contain four times more job transitions than ad-hoc market definitions.

We outline four stylized facts from the data-driven markets. First, there is substantial heterogeneity in market size and market concentration. The largest 50 markets contain around 55% of job transitions and have more than 100 nodes each. On the other hand, 40 thousand markets have a single node. Most markets have several firms and are not concentrated. We show that 86.7% of workers participate in labor markets that are not concentrated. This should not cast a shade on the fact that more than 10% of workers participate in highly concentrated markets. Data-driven labor markets are usually less concentrated than finer ad-hoc labor market (like county by 6-digit occupation) and more concentrated than broad ad-hoc market (like commuting zone by 3-digit industry).

Second, we show that firms hire workers from several different markets. While 76.4% of firms hire workers from a single market, the 13,308 largest firms—which employ 75% of formal workers in Brazil—hire workers from an average of 10.9 markets and a median of 7 markets. Firms participate in different markets either because they have establishments in many locations or because they hire workers from different occupations. Our results are complementary to the approach by Nimczik (2020). The author uses a similar method to define data-driven labor markets in Austria where nodes are constituted by single firms. As a consequence, in his paper firms cannot participate in different markets. We show that, because markets are spread across occupations, firms may participate in many markets.

Third, we characterize markets by their geographical dispersion. We show that workers participate in markets that are spread across the country (4%), across neighboring states (2%), and across microregions (10%). While 32% of workers participate in markets that include many counties within a microregion, for some workers this definition is too broad. In fact, 27% of workers participate in very local markets that do not cross county borders.

For last, we show that most markets include many occupations and industries. For example, 88% of markets have nodes in at least two different 3-digit occupations. In general, data-driven labor markets pass the common-sense test. For example, the relevant market for airline pilots is a detailed occupation-industry cell at the national level. On the other hand, the relevant market for salespeople consists of many occupations and industries but are restricted to a narrow geographic area.

This paper contributes to the growing literature on labor market definitions. Schmutte (2014) and Nimczik (2020) use different methods from the network literature to identify markets from worker flows (Modularity maximization and Stochastic Block Model). Both studies find that most job transitions are concentrated in very few markets. However, this could be a consequence of the lack detailed industry and occupation data. We use county by 6-digit occupation by 5-digit industry nodes, which allows us to better characterize job transitions. In addition, the CPM method and the algorithm used in this paper have the advantage of being able to identify markets with few nodes. Manning and Petrongolo (2017) estimate a spatial job search model and find that workers have strong preferences for jobs that are in close proximity. We show that while this can be true for most workers, individuals in some occupations and industries have job transition that cross long distances. Schubert et al (2021) take a different approach by creating a measure of outside-occupation options.

Our data-driven labor markets can be useful to guide policy. In particular, antitrust agencies should be careful when evaluating the potential effects of mergers and acquisitions. A result from this paper is that a merger between two large firms may affect workers in more than a single market and the effects could differ depending on each market’s concentration. Prager and Schmitt (2021) and Guanziroli (2022) do a retrospective merger analysis and show that the effects of the merger vary by occupation, consistent with our results.

2. Literature has different Local Labor Market Definitions

In this section, we analyze how recent studies from three different topics define labor markets. For the papers that do not explicitly define labor markets, we use the unit of observation from the main results of the paper.

Appendix Table 1 presents nine articles that study the effects of increasing labor market concentration on worker's wages and employment. Market definitions are especially important in this literature, as they matter for calculating the level of concentration. In all studies, market definitions seem to be dependent of data availability. Studies that use employer-employee data from the United States tend to define labor markets with the commuting zone X industry cell, but they disagree on the level of aggregation (Benmelech et al, 2020; Arnold, 2021; Berger et al, 2022). Markets are not defined using occupation codes due to the LEHD not having occupation information. That said, Azar et al (2020, 2022) gather occupation information from online job postings and argue that commuting zone X occupation is a better labor market definition. Brook et al (2021) and Marinescu et al. (2021) use employer employee data from India and France, respectively, and also define labor markets based on data availability (district X industry and Commuting zone X occupation). Prager and Schmitt (2021) and Guanziroli (2022) analyze the effects of mergers within an industry. Using detailed data, they are able to define markets at the location X industry X occupation level.

Appendix Table 2 shows that the minimum wage literature also has a disagreement regarding market definitions, which is probably due to different data availability. Under the monopsony theory, increases in minimum wage should not affect employment in highly concentrated markets. Hence, market definitions are relevant to these studies. That said, most studies in this literature implicitly use aggregate labor market definitions, like state X industry. As a consequence, the estimated minimum wage effects might be aggregating different heterogenous effects. In addition, there could be contamination between treated and control groups in some studies as markets often cross state boundaries.

Market definitions also vary across papers that study the effects of trade shocks on labor market outcomes. Recent studies show that trade shocks, like trade liberalization, affect some sectors of the economy more than others, a fact that has been explained by the presence of mobility frictions. Market definitions are important to these contexts since the effects of trade shocks are probably heterogenous by labor market. Appendix Table 3 shows that these definitions also vary across studies.

3. The Brazilian Matched Employer-Employee Dataset

To identify worker flows we use data from RAIS (Relação Annual de Informações Sociais), the Brazilian employer-employee linked administrative dataset.

RAIS is a confidential dataset maintained by the Brazilian Ministry of Labor. We use the data from 2007 to 2014. The dataset includes comprehensive information of firms, establishments, workers, and of the job match. The data is disclosed annually, and firms have to report all their job links within a year. Workers may have more than one entry in a single year, as they switch firms. In this paper, we will use information on occupation, industry sector, and county of employment. To identify worker flows we also use workers' and firms' identification numbers.

The method used in this paper requires the collection of job transition data. The RAIS dataset is well-suited for this purpose because it can provide a large number of transitions between very fine cells, or hereafter, nodes. In this paper, **a node is a county X 5-digit industry X 6-digit occupation cell**. We observe workers switching counties, industry, and occupations when they switch jobs. Thus, we can compute the number of switches for every pair of nodes. The link between two nodes is called an **edge**. Note that an edge may be constituted of one or more **job transitions** and the number of job transitions is referred to as the **weight** of an edge. In the data, job transitions come from: (i) transitions across consecutive years, and (ii) transitions within the year. We compute job transitions for the years between 2007 and 2013.

We impose the following restrictions to the data: First, we exclude workers from the public sector. The reason is that even though workers from the public sector work in very distinct jobs they are sorted into a single industry category. Although governments are an important participant of the labor market, their inclusion would probably agglomerate very distinct nodes, leading to misclassification. Secondly, we exclude temporary workers, internships, and contracts with less than 30 monthly hours of work. These workers are usually not relevant to firms' operation, but given their nature, they might be overrepresented in terms of job transition. Lastly, we exclude workers with more than two observations within a year, as these might reflect reporting mistakes and their inclusion could create false transitions between disconnected nodes.

Column 1 of Table 1 presents sample sizes of the resulting *raw data*. Between 2007 and 2013, we identified workers in 5,360,609 nodes and their transitions between nodes led to 45,270,686 edges. However, most nodes are too small, and most edges contain a single transition. Hence, to reduce noise and prevent the use of misclassified transitions, we impose additional restrictions. We exclude nodes—and all their edges—with weight lower than five. I.e., we exclude nodes in which less than five workers switched jobs between 2007 and 2013 to another firm in the same node. We also exclude edges with only one transition. Column 2 of Table 1 presents sample sizes of the *main data* used in the paper. We are left with 246,638 nodes and 6,819,564 edges. There is an average of 178 transitions per node, but many transitions occur within nodes (35% of them.). I.e., workers tend to switch jobs to firms in the same county, same occupation, and same industry.

Table 1: Job transitions and sample size

	Raw data	Main data
Nodes (county X 5-digit industry X 6-digit occupation)	5,360,609	246,638
Edges	45,270,686	6,819,564
Transitions (2007-2013)	82,321,061	43,874,702
Transitions between the same node (% of total transitions)	9.9%	34.8%
Average edges per node	8.4	27.7
Average transitions per node	15.4	177.9
Average transitions per node (except own)	13.8	115.9

Note: The table presents sample sizes of the raw dataset and the main dataset used in the paper. A node is a county X 5-digit industry X 6-digit occupation cell. Edges are non-empty links (or transitions) between nodes. Edges between two cells may have more than one transition. RAIS 2007-2013.

In the next section, I discuss a method that aggregates nodes into markets. Appendix Table 4 shows that, after data cleaning and restrictions, we are left with 4,246 counties, 2,026 6-digit occupations and 646 5-digit industry sectors. Their combination leads to 246,638 nodes. Studies using data with less precise categories or missing industry, occupational or regional data will have fewer nodes. Depending on the structure of job transitions, these could lead to few markets being detected or to wrongful detection.

4. A Method to Define Labor Market Approximations

This section describes the method to identify communities, which are approximations to labor markets. The Labor market is a complex network of workers and firms. While an individual entering the market can in theory apply for any available job, workers and firms tend to cluster at different levels of occupation, industry sector, and location. The goal of this paper is to identify such clusters, or communities.

To identify communities, we use a method called constant Potts model (CPM). The CPM method is a reiterative process that assigns nodes to communities until there are strong connections between the nodes within a community and weak connections between nodes across communities.

The notation is the following: the connected graph G includes n nodes and m edges. Nodes take the atomistic form and may also be referred to as vertex. Edges are the structure within a graph that attach two nodes. The adjacency matrix $A_{n\times n}$ maps the edges between nodes, where $A_{ij} = 1$ if there is an edge between nodes i and j. Edges have weight w_{ij}. The community of node i is denoted by σ_i and the function $\delta(\sigma_i, \sigma_j)$ takes value one if $\sigma_i = \sigma_j$, i.e., nodes i and j belong to the same community, and zero otherwise.

The constant Potts model maximizes the expression below by choosing a value σ_i for all i.

$$H = \sum_{ij} \left(A_{ij} w_{ij} - \gamma \right) \delta\left(\sigma_i , \sigma_j \right)$$

Where γ is a constant, also called the resolution parameter. Traag et al (2011) show that γ balances the trade-off between maximizing the number of internal edges within a community and keeping communities relatively small. The constant can also be understood as a penalty for the inclusion of a new edge in the community. As a result of this maximization, communities should have strong links within them and weak links across them.

An alternative method previously used in the literature is the modularity maximization (Schmutte, 2014; Nimczik, 2020). Similar to CPM, the modularity maximization approach yields communities with strongly connected nodes. However, Fortunato and Barthélemy (2007) show that modularity maximization has a problem named the resolution limit problem that prevents the detection of smaller communities. The CPM method is resolution-limit-free (Traag et al, 2011).

4.1. Algorithm and implementation

In this paper, nodes are defined by the intersection between 6-digit occupations, 5-digit industries and counties. This is the finest cell in the data. Two nodes are connected by an edge if at least two worker switched jobs between these nodes. The edge weight is the total number of job transitions between the nodes. Hence, the CPM method aggregates occupations, industries, and counties with dense worker transition into a community, or a data-driven labor market.

The CPM method is implemented through Python using the Leiden algorithm (Traag, Waltman and van Eck, 2019). The Leiden algorithm corrects a flaw in the widely used Louvain algorithm which generates badly connected communities. We set the resolution parameter in the CPM method to $\gamma = 0.1$.

5. Characterizing Labor Markets

The Leiden algorithm identifies 54,855 communities, or approximations to labor markets. Column 1 of Table 2 presents descriptive statistics of the data-driven markets. The table shows that 75% of job transitions occur within data-driven markets, with the rest occurring across markets. This is an indicator of the algorithm's success.

The algorithm aggregates nodes into markets to maximize the density of a market and minimize cross market density. As a comparison, Column 2 of Table 2 presents descriptive statistics of a commonly used ad-hoc labor market definition—microregion X 6-digit occupation cells— that gives a similar number of markets to the data-driven method. The table shows that 45.8% of job transitions occurred within the microregion X 6-digit occupation cells. To highlight the contrast, I exclude transitions within the same node: 61.8% of job transitions occurred within data-driven markets, while only 16.8% occurred within microregion X 6-digit occupation cells.

Next, I discuss four stylized facts from the analysis of data-driven labor markets.

(i) There is substantial heterogeneity in market size and market concentration

Table 2 reveals considerable heterogeneity in market size. Markets have on average 4.5 nodes and 800 job transitions. But nodes are not equally distributed across markets. While 40,171 markets have a single node, nine markets have more than 1000 nodes. A significant number of markets is in between, with 14,675 markets having between 2 and 1000 nodes.

This translates into two facts: (a) A great share of workers and transitions are part of few large labor markets, and (b) A considerable share of workers participates in small and concentrated labor markets. The top 50 markets with more transitions have 55% of the transitions in the data and a proportional fraction of workers. That said, 950 markets have 33% of transitions and 53,000 markets have 11.6% of transitions. These numbers contrast with Schmutte (2014) and Nimczik

(2020) that find that most transitions and workers belong to very few markets. This discrepancy could be either to differences in context or in the choice of algorithm and data availability.

Table 2: Descriptive Statistics of Markets

	Data-driven Markets	Microregion X 6-digit occ.
Number of markets	54,855	71,361
Number of nodes	246,638	
Number of transitions	43,874,702	
Transitions within markets / Total transitions	**75.1%**	**45.8%**
Transitions within markets / Total transitions (excludes transitions within the same node)	**61.8%**	16.8%
Nodes per market	4.5	3.5
Transitions per market	799.8	614.8
Markets with…		
1 node	40,171	39,254
2 to 5 nodes	6,912	23,672
6 to 9 nodes	3,800	4,021
10 to 99 nodes	3,666	4,303
100 to 999 nodes	297	111
1000 or more nodes	9	0
Maximum number of nodes in a market	2,794	478
Share of job transitions in the…		
Top 50 markets	55.4%	17.0%
Top 100 markets	64.6%	23.7%
Top 500 markets	83.4%	44.5%
Top 1000 markets	88.4%	55.2%

Note: Nodes are defined as the county X 6-digit occupation X 5-digit industry cell. The first column presents descriptive statistics of the data-driven labor markets that are produced through the Leiden algorithm. The second column presents descriptive statistics of a commonly used ad-hoc labor market definition where a market is defined by the intersection between a microregion and a 6-digit occupation. RAIS 2007-2013

Panel A of Table 3 shows that some markets are highly concentrated, and some are not. Labor market concentration is measured with the Herfindahl–Hirschman Index (HHI). The HHI for a single market is defined as the sum of the square of firm's employment share in that market. The scale ranges between 0 and 10,000, with 0 representing perfectly competitive markets and 10,000 representing a monopoly. The average HHI across all markets is of 6,450.7 points, which is substantially high. To put into context, the U.S Department of Justice (DOJ) considers that product markets with HHI between 1,000 and 1,800 are moderately concentrated and markets with HHI above 1,800 are highly concentrated.

However, when considering the size of each market, we learn that most workers participate into markets that are not concentrated. The previous numbers are led by markets of a single node and that have very few workers. The average HHI weighted by employment in 2014 is of 614.6. Appendix Table 5 shows that 86.7% of workers participate in labor markets with HHI lower than 1000 points, which are not concentrated according to the DOJ's guidelines.

The use of ad-hoc proxies for labor markets, such as the microregion X 6-digit occupation definition leads to the misleading conclusion that markets are highly concentrated. Table 3 shows that the average HHI weighted by employment is of 1,341 points. In addition, the metrics suggests that more than 25% of workers participate in moderate to high concentrated markets. However, ad-hoc proxies fail to aggregate workers and firms by how interconnected they are. For example, this measure allocates airplane pilots into different regional markets, each with very few firms. This would lead to a highly concentrated market even though there are many firms at the national level. The next stylized facts show that markets often cross regional boundaries.

(ii) Firms hire workers from many markets

Most firms hire workers from a single market, but the largest firms hire workers from many labor markets. Panel B of Table 3 shows the distribution of firms by the number of markets in which they hire workers. In 2014, we identified 1.4 million firms, and more than 75% of them hire workers from a single market. However, the 13,308 largest firms (by labor force size) hire workers from an average of 10.9 markets. The largest 738 firms in the country hire workers from an average of 40.9 markets.

Labor market definitions are important when considering the effects of mergers and acquisitions. A merger between two firms increases concentration in the labor market since there is one less firm in the market. Measurement of increases in concentration depend on the market definition. A merger in a market with 1,000 equally sized firms will not significantly increase concentration, but a merger in a market with 10 firms will. The increase in concentration (measured by the HHI) is one of the main metrics used by antitrust agencies around the world when analyzing product markets. Our approach shows that firms hire workers from many markets. As a consequence, they may increase concentration in some of these markets, but not necessarily in all markets.

Prager and Schmitt (2021) and Guanziroli (2022) show the importance of labor market definitions when analyzing the effect of a merger in the hospital and retail pharmacy sectors, respectively. In both their analysis, firms are assumed to hire workers from different markets, as defined by their occupations. The authors show that this margin is important, with the effects of mergers being different across occupations. Our approach provides a guide to future researchers and antitrust agencies on how to determine the markets in which firms participate.

Table 3: Labor market concentration and number of markets by firm

	Min	p25	p50	Mean	p75	Max	N
Panel A: HHI							
Data-driven markets							
Market distribution	9.7	2800.0	7222.2	6450.7	10000	10000	54,855
Worker distribution	9.7	17.8	51.2	614.6	266.6	10000	16,828,223
Microregion X 6-digit occupation							
Market distribution	3.2	1533.3	4583.3	5194.3	10000	10,000	71,361
Worker distribution	3.2	59.5	249.6	1341.3	1136.6	10000	16,828,223
Panel B: Number of markets per firm							
All firms	1	1	1	1.5	1	757	1,440,839
Top employers (75% of workers)	1	3	7	10.8	13	757	13,308

Top employers (50% of workers)	1	9	22	40.9	47	757	738

Note: Panel A presents the distribution of markets over the Herfindahl Hirschman Index (HHI). The HHI was computed using employment shares of each firm in the market. Panel B presents the distribution of firms by number of markets in which they hire workers from. The line "Top employers (75% of workers)" refer to the firms, ranked by employment size, that employed 75% of the workers in Brazil in 2014. Data: RAIS 2014.

(iii) Some markets are geographically dispersed

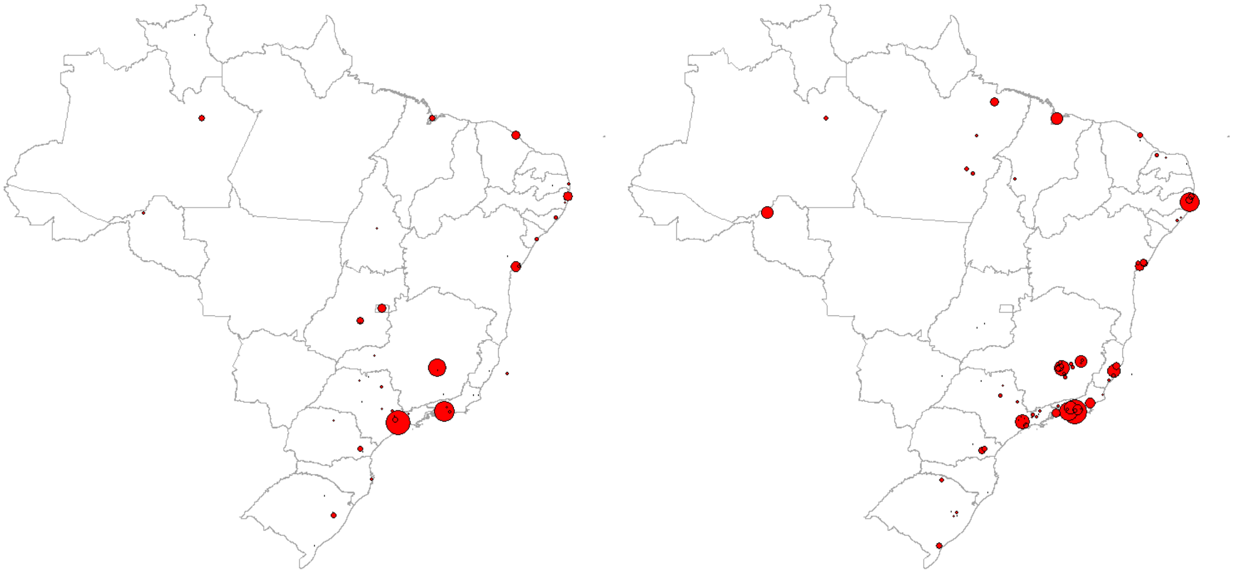
One of the main problems of using ad-hoc labor market proxies such as occupation X microregion cells is that some types of workers move across regions. In our approach, we find that some workers participate in markets that are spread across the country, across neighboring states, across microregions and within a state. On the other hand, some markets are smaller than a commuting zone, or a microregion. Next, we discuss each case.

a. Nationally

Figures 1.A and 1.B present two examples in which the data-driven markets are spread across the country. Figure 1.A shows market number 91, which is composed mostly of workers in the civil engineering occupation. In the figure, each circle represents a county that belongs to the market. The size of the circle represents the sum of transitions of all nodes that belong to market 91 in that county. Market 91 is composed of 233 nodes that include 56 counties, 21 occupations and 29 industries. Civil engineers cover 59% of nodes and the other occupations are closely related even though some are categorized in different 1-digit occupation. The market is defined over 53,832 transitions.

Nationally dispersed labor markets are not necessarily associated to workers with a college degree. Figure 1.B shows market number 80, which is composed mostly of workers in the welder 6-digit occupation. Market 80 is composed of 252 nodes that include 82 counties, 13 occupations and 51 industries. The welder occupation is in 84.1% of nodes. The market is defined over 64,959 transitions.

Table 4 shows that many workers participate in nationally dispersed markets.[1] There are 432 markets classified as nationally dispersed and they contain 4.8% of job transitions. Interestingly, nationally dispersion is not a characteristic of all large markets. From the 1000 largest markets in terms of transitions, only 39 are nationally dispersed.



Panel A: Market #91
Modal occupation: Civil Engineer

Panel B: Market #80
Modal occupation: Welder

Figure 1. Nationally dispersed markets

Note: The figures describe two distinct data-driven labor markets. Each circle represents a county from a node in the market, and the size of the circle is proportional to the number of workers from that county that participate in the market.

b. Across Neighboring States

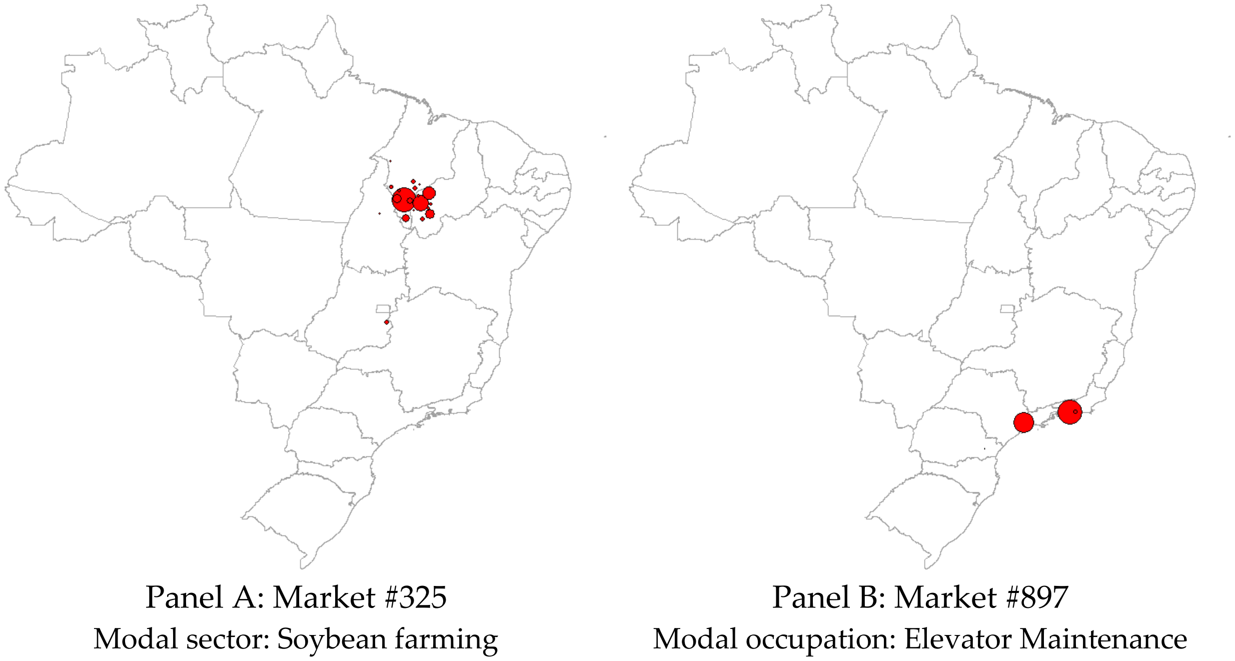
Some markets cross state borders but are not nationally dispersed. Figure 2 plots the map of markets 325 and 977. Market 325 has soybean farming as the modal sector, containing 68% of job

[1] I define nationally dispersed labor markets the markets that have nodes in more than one region and in more than two states and satisfy at least one of the following conditions: (1) more than 15% of transitions in at least two different regions; (2) more than 10% of transitions in at least three regions; (3) more than 5% of transitions in at least four regions; (4) more than 2.5% of transitions in all regions; (5) more than 5% of transitions in at least four states; (6) more than 2.5% of transitions in at least six states; and (7) more than 1.5% of transitions in at least eight states.

transitions. The modal occupation is tractor driver, containing only 15% of transitions. Most of the other occupations are in the context of operating agricultural machines or agriculture work. Panel A of Figure 2 shows that most nodes of market 325 are contiguous and concentrated in the states of Maranhão and Piauí.

Panel B of Figure 2 shows market 897, which is contained in the states of São Paulo and Rio de Janeiro, with most transitions occurring in the capitals of these states. The market includes 20 occupations and 9 industries, with elevator maintenance having only 20.5% of transitions. Other occupations are within the scope of electromechanics.

Table 4 shows that there are 401 markets that cross neighboring states, containing only 2.1% of job transitions.


Panel A: Market #325
Modal sector: Soybean farming

Panel B: Market #897
Modal occupation: Elevator Maintenance

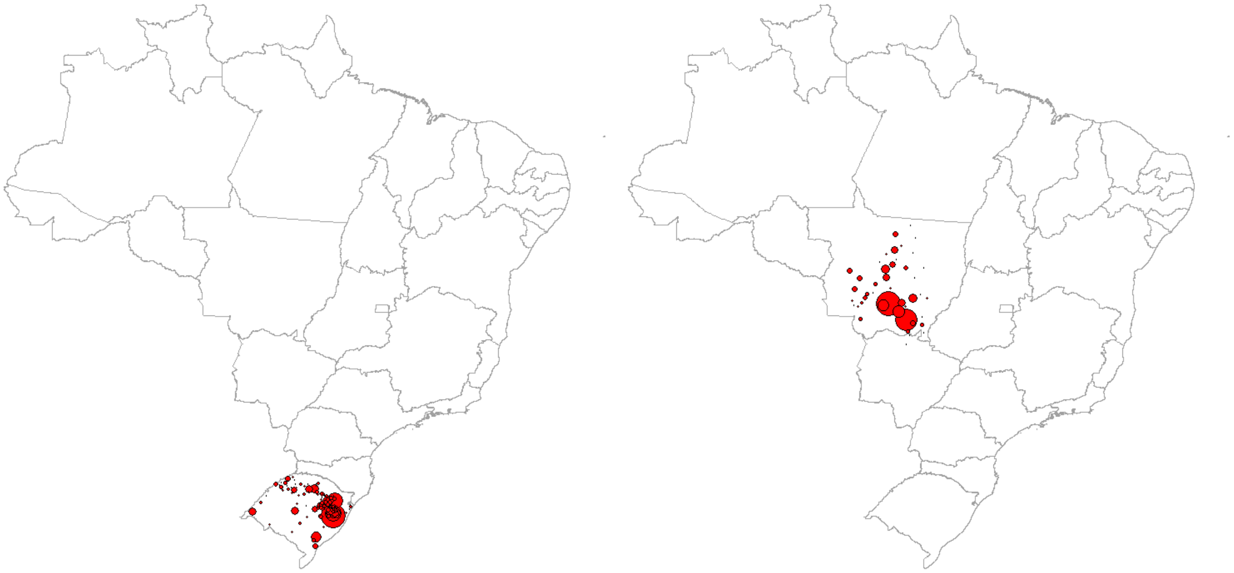
Figure 2. Markets in neighboring states

Note: See Figure 1.

c. Within States, across microregions

Some markets cross commuting zones, or microregions, but do not cross state boundaries. Figure 3 plots the nodes of markets 49 and 133. The modal occupation in both is truck driver (modes have 66% and 77% of transitions, respectively), but market 49 is restricted to the state of Rio Grande do Sul and market 133 is restricted to the state of Mato Grosso. Each market includes nodes from around 50 industries.

A substantial number of workers participate in state dispersed markets. Table 4 shows that 1,194 markets are state dispersed, containing 10.6% of job transitions.



Panel A: Market #49
Modal occupation: Truck Driver

Panel B: Market #133
Modal occupation: Truck Driver

Figure 3. State dispersed markets

Note: See Figure 1.

d. Within microregions

Most workers participate in markets that are concentrated within a microregion and include many counties. Table 4 shows that there are 2,830 markets—containing 32.2% of job transitions—in which more than 90% of transitions occurred within a single microregion (and are not part of the next category). From the top 1000 markets, 266 are contained within a microregion.

e. Within counties

Most markets pertain to a single county. Table 4 shows that there are 46,059 markets in which more than 90% of job transitions occurred in a single county. These markets correspond to 27% of job transitions, and while most markets are small and have a single node, 369 markets are within the top 1000 markets.

A substantial number of markets does not fit in any category. Most of these markets are concentrated within a county but have a reasonable share of transitions in other counties, microregions or states.

Table 4: Types of markets by geography

Market type	All		Top 1000 markets	
	Count	Transitions (%)	Count	Transitions (%)
Panel A: By Geography				
a. Nationally dispersed	432	4.8%	39	4.3%
b. Neighboring states	401	2.1%	17	1.9%
c. State dispersed	1,194	10.6%	149	9.4%
d. Single Microregion	2,830	32.2%	266	29.8%
e. Single county	46,059	27.0%	369	22.2%
f. Other categories	3,622	23.4%	160	32.5%

Note: See footnote 1 for definition of nationally dispersed markets. Top 1000 markets are ranked by number of transitions.

(iv) Most markets include many occupations and industries

Data-driven markets are incredibly sparse across industries and occupations. Table 5 shows that most markets have nodes in two or more different 2-digit industries and 1-digit occupations (69% and 78% of markets, respectively). While ad-hoc labor market definition yields similar numbers for industry dispersion, by construction, these markets are not spread across occupations.

Table 5: Market composition

Share of markets composed of at least two different…	Data-driven Markets	Microregion X 6-digit occ.
States	26%	-
Mesoregions	43%	-
Microregions	53%	-
Counties	69%	69%
2-digit industries	69%	70%
3-digit industries	76%	81%
1-digit occupations	78%	-
2-digit occupations	83%	-
3-digit occupations	88%	-
Number of markets	14,684	32,107

Note: The table only includes markets that have more than one node.

6. Discussion

The labor market is a global network of firms and individuals in which firms offer wages to individuals in order to compensate them for their work. All workers and firms are somehow connected, such that a shock in the soybean production in Argentina will eventually affect the wages of workers in the retail sector in China. That said, workers tend to cluster by some observable characteristics, like occupation, industry, and geographic location. Any event within the cluster should affect workers and firms in a faster and more direct way.

In this paper, we presented an empirical method that attempts to identify such clusters by using job transitions across fine occupation-industry-location cells. We refer to clusters as data-driven labor markets, or labor market approximations. The method used in this paper—the constant

Potts model—is successful in identifying data-driven labor markets where most job transitions occur within labor markets, and not across labor markets.

There are four important takeaways from the analysis of data-driven labor markets: (i) There is substantial heterogeneity in market size and market concentration; (ii) Firms hire workers from many different markets; (iii) Some markets are geographically dispersed; and (iv) Most markets include many occupations and industries. These takeaways highlight the benefits of using the labor markets identified through our method instead of ad-hoc labor market definitions.

In this paper we did not verify the robustness and validity of data-driven labor markets. First, the CPM method requires the choice of a constant, which will help determine the number of markets. It is possible that changes in this constant will change the composition and number of markets. Future work should verify the robustness of data-driven labor markets to changes in the Potts constant. Second, it is necessary to develop an approach that determines if data-driven markets predict the consequences of events in the real world. For example, do movements in wages correlate across workers within markets? Does the effect of a plant closure on workers wages dissipate as predicted by the data-driven labor markets? And do increases in minimum wages reduce employment of workers in less concentrated markets? Future work should check the validity of data-driven labor markets.

References

Adão, Rodrigo. 2016. “Worker Heterogeneity, Wage Inequality, and International Trade: Theory and Evidence from Brazil.” Unpublished.

Arnold, David. 2021. “Mergers and Acquisitions, Local Labor Market Concentration, and Worker Outcomes”. Unpublished.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." American Economic Review, 103 (6): 2121-68.

Azar, Jose, Emiliano Huet-Vaughn, Ioana Marinescu, Bledi Taska, and Till von Wachter. 2019. “Minimum Wage Employment Effects and Labor Market Concentration.” NBER Working Papers 26101. Doi:10.3386/w26101

Azar, Jose, Ioana Marinescu, Marshall I. Steinbaum, and Bledi Taska. 2020. “Concentration of US Labor Markets: Evidence from Online Vacancy Data.” Labour Economics 66: 101866.

Azar, Jose, Ioana Marinescu, and Marshall I. Steinbaum. 2022. “Labor Market Concentration.” Journal of Human Resources. DOI:10.3368/jhr.monopsony.1218-9914R1.

Benmelech, Efriam, Nittai Bergman, and Hyunseob Kim. 2020. “Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?” Journal of Human Resources. DOI:10.3368/jhr.monopsony.0119-10007R1.

Berger, David W., Kyle Herkenhoff, and Simon Mongey. 2022. “Labor Market Power.” American Economic Review. Vol 112(4):1147-93. DOI: 10.1257/aer.20191521

Brooks, Wyatt J., Joseph P. Kaboski, Illenin O. Kondo, Yao Amber Li and Wei Qian. 2021. “Infrastructure Investment and Labor Monopsony Power.” IMF Economic Review, 69(3), 470-504.

Caliendo, Lorenzo, Maximiliano Dvorkin and Fernando Parro. 2019. “Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock.” Econometrica. Vol: 87 (3): 741-835. https://doi.org/10.3982/ECTA13758

Cengiz, Doruk, Arindrajit Dube, Attila Lindner and Ben Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs.” The Quarterly Journal of Economics. Volume 134, Issue 3, August 2019, Pages 1405–1454, https://doi.org/10.1093/qje/qjz014

Dix-Carneiro, Rafael, and Brian K. Kovak. 2017. "Trade Liberalization and Regional Dynamics." American Economic Review, 107 (10): 2908-46.

Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. "Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties." The Review of Economics and Statistics 92, no. 4: 945–64.

Dustmann, Christian, Attila Lindner, Uta Schönberg, Matthias Umkehrer and Philipp vom Berge. 2022. "Reallocation Effects of the Minimum Wage". The Quarterly Journal of Economics, Volume 137, Issue 1, February 2022, Pages 267–328, https://doi.org/10.1093/qje/qjab028

Felix, Mayara. "Trade, Labor Market Concentration, and Wages". 2022. Unpublished.

Fortunato, Santo, and Marc Barthélemy. 2007. "Resolution Limit in Community Detection." Proc. Natl. Acad. Sci. U. S. A. 104 (1): 36–41. http://dx.doi.org/10.1073/pnas.0605965104.

Guanziroli, Tomás. 2022. "Does Labor Market Concentration Decrease Wages? Evidence from a Retail Pharmacy Merger." Unpublished.

Hakobyan, Shushanik, and John McLaren. 2016. "Looking for Local Labor Market Effects of NAFTA." The Review of Economics and Statistics; 98 (4): 728–741.
doi: https://doi.org/10.1162/REST_a_00587

Kovak, Brian K. 2013. "Regional Effects of Trade Reform: What Is the Correct Measure of Liberalization?" American Economic Review, 103 (5): 1960-76.DOI: 10.1257/aer.103.5.1960

Manning, Alan, and Barbara Petrongolo. 2017. "How Local Are Labor Markets? Evidence from a Spatial Job Search Model." American Economic Review, 107 (10): 2877-2907. DOI:10.1257/aer.20131026

Marinescu, Ioana, Ivan Ouss and Louis-Daniel Pape. 2021. "Wages, hires, and labor market concentration." Journal of Economic Behavior & Organization. Vol 184: 506-605.

Nimczik, Jan Sebastian. 2020. "Job mobility networks and Data-driven Labor Markets."

Prager, Elena, and Matt Schmitt. 2021. "Employer Consolidation and Wages: Evidence from Hospitals." American Economic Review, 111 (2): 397-427.

Saltiel, Fernando Andres, and Sergio Urzua. 2022. "Does an Increasing Minimum Wage Reduce Formal Sector Employment? Evidence from Brazil". Economic Development and Cultural Change. Forthcoming.

Schmutte, Ian M. 2014. “Free to move? A network analytic approach for learning the limits to job mobility.” Labour Economics, 29: 49–61.

Schubert, Gregor, Anna Stansbury, and Bledi Taska. 2021. “Employer concentration and outside options.” Available at SSRN 3599454.

Traag, V. A., Van Dooren, P., Nesterov, Y. 2011. Narrow scope for resolution-limit-free community detection. Phys. Rev. E, 84:016114. 10.1103/PhysRevE.84.016114

Traag, V.A., Waltman. L., Van Eck, N.-J. (2018). From Louvain to Leiden: guaranteeing well-connected communities. arXiv:1810.08473

Appendix Tables

Appendix Table 1: Recent Literature on Labor Market Concentration

Paper	Market Definition	Country
Azar et al. (2020)	Commuting zone X 6-digit occupation	US
Benmelech et al. (2020)	County X 4-digit industry Commuting zone X 4-digit industry	US
Prager and Schmitt (2021)	Commuting zone X Hospitals X Occupation Commuting zone X All health care X Occupation	US
Arnold (2021)	Commuting zone X 4-digit industry	US
Azar et al. (2022)	Commuting zone X 6-digit occupation	US
Berger et al. (2022)	Commuting zone X 3-digit industry	US
Brooks et al. (2021)	District X 4-digit industry State X 4-digit industry	India
Marinescu et al. (2021)	Commuting zone X 4-digit occupation X quarter	France
Guanziroli (2022)	County X Retail Pharmacies X Occupation	Brazil

Note: The table presents a selected sample of recent studies in the labor market concentration literature. Colum 2 describes the market definition explicitly or implicitly used in each study.

Appendix Table 2: Recent Literature on the Effects of Minimum Wages

Paper	Market Definition	Country
Saltiel and Urzua (2022)	Microregion	Brazil
Dube, Lester and Reich (2010)	County X Restaurants	US
Cengiz et al. (2019)	State X Demographic groups State X 1-digit industry	US
Dustman et al (2022)	District X Demographic groups	Germany
Azar et al. (2019)	County X 6-digit occupations	US

Note: The table presents a selected sample of recent studies in the minimum wage literature. Colum 2 describes the market definition explicitly or implicitly used in each study.

Appendix Table 3: Recent Literature on the Effects of Trade Shocks

Paper	Market Definition	Country
Felix (2022)	Microregion X 6-digit occupation	Brazil
Adao (2016)	Microregion X Schooling	Brazil
Kovak (2013)	Microregion	Brazil
Dix Carneiro and Kovak (2017)	Microregion	Brazil
Autor, Dorn and Hanson (2013)	Commuting zones X SIC codes	US
Hakobyan and McLaren (2016)	County X 2-digit industry	US
Caliendo, Dvorkin and Parro (2019)	States X 12 manufacturing sectors	US

Note: The table presents a selected sample of recent studies in the empirical trade literature. Colum 2 describes the market definition explicitly or implicitly used in each study.

Appendix Table 4: Sample size

	Raw data	Main Data
Region:		
States	27	27
Microregions	137	137
Mesoregions	558	553
Counties	5551	4246
Occupation:		
1-digit	10	10
2-digits	49	45
3-digits	193	185
4-digits	614	577
6-digits	2588	2026
Industry:		
2-digits	88	87
3-digits	285	274
5-digits	673	646

Note: The table presents the number of regions, occupations and industries in the raw dataset and in the main dataset used in the paper. RAIS 2007-2013.

Appendix Table 5: Labor market concentration

Data-driven markets	# markets	# workers	% workers
HHI <=1000	6,033	14,598,459	86.7%
1000>HHI>=1800	3,466	527,595	3.1%
HHI>1800	45,356	1,702,170	10.1%

Note: The table categorizes markets in terms of concentration levels, as measured by the Herfindahl–Hirschman Index (HHI). The number of workers comes from RAIS 2014.