Detecting Person Identifier Misuse: Understanding the Longitudinal Employer Household Dynamics (LEHD) Pseudo-PIK Process

Kevin L McKinney (Kevin.L.McKinney@census.gov)

and

Stephen R Tibbets (Stephen.R.Tibbets@census.gov)

April 10, 2026

Abstract: The Longitudinal Employer Household Dynamics (LEHD) program at the U.S. Census Bureau uses primarily administrative data to release public use statistics such as the Quarterly Workforce Indicators (QWI). Administrative data from state unemployment insurance (UI) systems provides extensive coverage of quarterly job earnings, however the person identifier attached to UI earnings reports is susceptible to misuse. The Pseudo-PIK process was developed to detect work histories where multiple people use the same Person Identification Key (PIK). The pseudo-PIK process assigns a new person identifier to each selected job along with a complete set of imputed demographic characteristics. The pseudo-PIK process improves the quality of tabulations by person demographic characteristics, while not affecting tabulations by firm characteristics. For example, in some NAICS 4 industries such as "1113: Fruit and Tree Nut Farming" a relatively large share of job quarters are assigned a pseudo-PIK, greatly increasing the share of Hispanic workers while leaving the total number of jobs in the industry unchanged. The ultimate impact of the Pseudo-PIK process on LEHD demographic tabulations depends primarily on the share of jobs with a pseudo-PIK and the distribution of their imputed characteristics. The higher the share of jobs with a pseudo-PIK and/or the extent pseudo-PIK imputed characteristics differ from the characteristics of the original UI reported PIK the larger the impact.

Introduction

The Longitudinal Employer Household Dynamics (LEHD) program at the U.S. Census Bureau produces several products such as the Quarterly Workforce Indicators (QWI) that rely on accurate employment histories. LEHD builds a worker's employment history from data collected to administer state unemployment insurance (UI) systems. A state's UI data consists of quarterly earnings reports for PIK (the internal LEHD person identifier) and state employer identification number (SEIN) combinations or what are commonly known as jobs.[1] However, the PIKs attached to reported earnings histories are not validated either by LEHD or the state that originally received them. If all workers consistently use the same nine-character identifier across all UI covered employers, then the LEHD PIK earnings histories should be complete subject to errors in reporting by the firm and errors in processing by state unemployment insurance systems. However, in practice reporting and processing is not perfect and some jobs are not assigned to the correct person. This is especially likely if a PIK has an unusually large number of jobs or jobs active during periods that conflict with other administrative data. The goal of the Pseudo-PIK process is to identify these types of work histories and assign each suspect job a unique nine-character PIK with all demographic characteristics imputed.[2]

If a worker is legally entitled to a valid government issued person identifier such as a Social Security Number (SSN), there is little incentive to present an alternative identifier when starting a new job. However, for persons without legal access to the US labor market, there are strong incentives to present a self-issued SSN as if it were a valid government identifier. There are three possibilities when choosing a nine-digit identifier; the nine-digit number has already been issued to another person by the Social Security Administration (SSA), the nine-digit number will never be issued by SSA, or the number has yet to be issued. If the PIK has already been issued then at least some of the jobs are likely to have been generated by the original SSN holder, however if the PIK is not in scope or has not been issued then we have no information on the demographic characteristics of any of the workers associated with the PIK.

Unfortunately, we have no training datasets available to develop models that identify PIKs used by multiple persons. Instead, we use two main strategies to detect work histories used by multiple workers. The first approach identifies unusual job histories using information from the work history itself, such as the total number of jobs. For the second approach we take advantage of administrative information from SSA to find periods of unexpected jobholding, for example jobs active prior to a person becoming legally eligible to work or jobs active after a person's death.

All PIK's created by the Pseudo-PIK process have only one job, the process does not combine suspect jobs into synthetic work histories. Combining the jobs into synthetic work histories would

[1] Internally, LEHD uses both the PIK SEIN and the PIK FID to identify jobs. We use the FID as the firm identifier for the Pseudo-PIK process. The FID is created using a separate process that identifies linked SEINs based on worker flows and QCEW reported SEIN-to-SEIN relationships. Please see McKinney and Tibbets (2025) for more information.

[2] The Pseudo-PIK process assigns identified jobs a new PIK starting with a "P". By design, the new PIKs are in an unused region of the PIK character space and will not match to external demographic data sources, triggering an imputation of all characteristics on the LEHD Individual Characteristics File (ICF).

require significant additional development resources with benefits unequally distributed across LEHD products. For example, not creating synthetic work histories has almost no impact on job-based products such as the QWI, but assigning each new PIK only one job will impact LEHD products such as Job to Job Flows (J2J) that rely on job transitions. The Pseudo-PIK process reduces spurious job-to-job transitions in J2J instead assigning these transitions (possibly incorrectly) to non-employment. For tabulations by demographic characteristics, imputations more accurately represent the characteristic uncertainty associated with pseudo-PIK jobs. Instead of assuming all jobs for a selected PIK have the same characteristics as the original SSN holder, each pseudo-PIK receives ten implicates or draws from the posterior predictive distribution for each characteristic.

PIK Types

The U.S. Census Bureau receives UI data from LEHD partner states with a 9-character person identifier attached to each quarterly earnings record. The person identifiers are generally 9-digit numbers, but in some cases may include other ASCII characters. The data from each state is received by the Census Bureau's Economic Reimbursable Surveys Division (ERD), where each person identifier goes through the "QPIK" process. A brief overview of the ERD QPIK process is shown below in Figure 1.

Figure 1: Person Identifier QPIK Process Overview

Stage	Action	PIK	SSN Flag
1	ID variable is compressed, removing null characters		
2	ID variable is not null, but contains non-numeric characters	000000000	C
3	Left pad ID variable with zeros if length is less than 9		
4	First three digits are not "000" or "666", the first digit is not 9, digits 4 and 5 are not "00", and digits 6 to 9 are not "0000"	QPIK lookup	V
5	ID variable is blank or all zeros	000000000	0
6	ID variables is all nines	999999999	N
7	ID variable starts with "000" but the other digits are not all zeros	QPIK "AA" lookup	A
8	Otherwise	QPIK lookup	I

The QPIK process assigns a PIK and SSN flag value to each record in a state's quarterly UI files. The files are then delivered to LEHD where they are known internally as the "uipik" files. The "uipik" files are the primary inputs used to construct the LEHD Employment History File (EHF).[3] The Pseudo-PIK process selects PIK values of "000000000", "999999999", and those starting with "AA" for special handling. We assume that PIK records with these values, which we henceforth call "AA" records, are not a valid person identifier by simple inspection and therefore should not reliably be able to be

[3] Prior to 2018 the state UI data was received and processed directly by LEHD. Each record was assigned a slightly different set of SSN flag values, although in either case the generated PIKs are the same.

used across employers. Every job (PIK FID[4] combination) for these records is assigned a unique pseudo-PIK value. For example, if a PIK with a value of "AA0000001" has 10 jobs, each job would receive a new pseudo-PIK. The records with an SSN flag value of "I" are not selected for special processing. The "I" flags primarily consist of records starting with a "9", which is the valid range for Individual Taxpayer Identification Numbers (ITIN) assigned by the IRS. Although an ITIN does not confer upon the recipient the legal right to work, it is a valid person identifier.

As part of a data sharing agreement with SSA, the US Census Bureau regularly receives the "SSA Numident". The SSA Numident contains all transactions for every legally issued SSN, from the first transaction when the SSN was issued to the last transaction when the person is reported dead. The Census Bureau transforms the SSA transaction data, creating a new file named the "Census Numident" containing one record per PIK. In addition to the SSN, the Census Numident contains information such as the date of first contact with SSA, date of birth, sex, place of birth, and date of death. Any PIK that matches to the Census Numident is almost always eligible to work in the U.S although there are some exceptions such as for student visas and temporary work authorizations.

Every PIK in the state UI data is matched to the Census Numident to determine if the PIK is legally eligible to work. A PIK that does not match to the Census Numident may have the legal right to work, but we do not have detailed information for these records. Table 1 below shows the distribution of records by PIK type for workers in the 2022 LEHD snapshot.

Table 1: Count and Share of Workers and Jobs by PIK type (disclosure pending DRB approval)

Table 1: Count and Share of Workers and Jobs by PIK type						
CNUM	Not on CNUM	Not on CNUM (AA)	Workers (PIK)	Share	Jobs (PIK FID)	Share
1	0	0	XXX,XXX,XXX	XX.XX	X,XXX,XXX,XXX	XX.XX
0	1	0	XX,XXX,XXX	X.XX	XX,XXX,XXX	X.XX
0	1	1	XXX,XXX	X.XX	XX,XXX,XXX	X.XX
		Total	XXX,XXX,XXX		X,XXX,XXX,XXX	

The overwhelming majority of workers and jobs have PIKs that match with a record on the Census Numident (CNUM in Table 1). For records that do not match, we group them into two categories; those that are not in the correct SSN format (Not on CNUM (AA)) and those that may be in the correct format (i.e. ITINs). Comparing the Workers and the Jobs columns provides information on the differences across PIK types in the average number of jobs per PIK. PIKs that match to the Census Numident and "AA" PIKs not on the Census Numident tend to have more jobs on average than PIKs in the "Not on CNUM" group alone. We will discuss this further in the next section.

[4] A FID is a within-PIK identifier that defines a job by linking the employment history SEINs across recognized successor-predecessor relationships. This is the primary unit of analysis for job-based LEHD data products.

PIKs with Many Jobs

Multiple persons using the same PIK observationally manifests itself when a PIK has an unusually large number of jobs over its entire history and/or during specific time periods.[5] PIKs with large numbers of jobs can noticeably impact LEHD's released statistics, especially when many of the workers in a firm, industry, or geographic area have PIKs with unrepresentative characteristics. For example, if large numbers of workers employed in the food services industry in Taos New Mexico use a small number of PIKs issued to older white males this will likely create significant bias in LEHD QWI published statistics for this geographic area tabulated by person characteristics such as race and education.

Insert Figure 2: CDF of PIKs and PIK FIDs by the Number of Jobs per PIK (disclosure pending DRB approval)

In Figure 2 we show the Cumulative Distribution Function (CDF) for both PIKs and PIK FIDs by the number of jobs per PIK. We show the CDF for both PIKs and PIK FIDs since LEHD primarily tabulates jobs (PIK FIDs) and PIKs with large numbers of jobs will have an outsize impact on our published statistics even though the number of PIKs with 30 or more jobs may be relatively small.

The overwhelming majority of PIKs (XX%) have 10 jobs or less over their observed work history. Also, only about X.X% of PIKs have 30 or more jobs, perhaps suggesting that multiple persons using the same PIK has relatively little impact on LEHD published statistics. However, the PIK level results are misleading compared with the PIK FID or job level results; workers with 10 jobs or less make up only XX% of all jobs, while PIKs with 30 or more jobs make up about XX% of all jobs. The PIKs with many jobs (30 or more) are almost 6 times larger than would be expected using only the distribution of the number of jobs per PIK.

Given the potentially large impact of PIKs with many jobs on LEHD statistics we first need to understand how the number of jobs differs across our three main PIK populations. In Figure 3A we show the CDF of PIKs by the number of jobs per PIK for each of the three populations (rows) shown in Table 1.

Insert Figure 3A: CDF of PIKs by Number of Jobs per PIK (disclosure pending DRB approval)

The "CNUM" line in Figure 3A is the reference group and is identical to the "Workers" line shown in Figure 2. Compared with the reference group, the "Not on CNUM" population has a very different

[5] Without using additional information such as the Census Numident, it is very difficult to identify PIK misuse when there are only one or two jobs in a quarter.

number of jobs distribution. Most of this population is "pseudo-PIK like"; XX% of the PIKs have only one job and since the records do not match to the Census Numident all person characteristics are imputed. The "Not on CNUM (AA)" distribution of the number of jobs per PIK is different (bimodal), with XX% of workers with only one job and XX% of workers with 30 or more jobs. However, as we show in Figure 3B this result is somewhat misleading. Figure 3B is like Figure 3A, except here we show the CDF of PIK FID's (instead of PIKs) by the number of jobs per PIK, which is more relevant for the typical LEHD tabulation.

Insert Figure 3B: CDF of PIK FIDs by Number of Jobs per PIK (disclosure pending DRB approval)

Once again, "CNUM" is the reference group and when comparing the PIK FID distributions striking differences emerge. The "Not on CNUM" group has a relatively large share of jobs associated with PIKs with only a single employer, XX% compared with X% for the reference group. This result is consistent with the results shown in Table 1 where the share of jobs in the "Not on CNUM" category is only X.XX% vs X.XX% for the share of workers. However, at the opposite end of the distribution both the CNUM and the "Not on CNUM" groups have a somewhat large share of PIK FID's associated with PIKs with 30 or more jobs, XX% and X% respectively. Although these results appear large in isolation, compared with the "Not on CNUM (AA)" group the share of jobs associated with PIKs with 30+ jobs is small. The "Not on CNUM (AA)" group share of jobs associated with PIKs with 30+ jobs is over XX%. Given that the SSN's underlying the AA group PIKs should be easy for the typical employer to detect we do not process this group through the Pseudo-PIK detection part of the process. We assume that every FID for a PIK in the AA group is used by a different person and assign every PIK FID in the AA group a pseudo-PIK.

PIK Pre-Selection

In the first step of the Pseudo-PIK process, we set initial criteria that are used to reduce the number of PIKs selected for further analysis. The pre-selection criteria attempt to capture all the PIKs or "record candidates" who would be identified for Pseudo-PIK assignment using the full sample, while greatly reducing the number of PIKs subject to intensive analysis. For the "too many jobs" criteria, PIKs must have at least 20 distinct employers (PIK FID), at least one quarter with 3 or more new jobs, and at least one quarter with 7 or more jobs. PIKs with more than 800 employers (PIK FID) are automatically selected to receive a pseudo-PIK.

Next, we match all the PIKs on the EHF with the Census Numident. This serves two purposes. First, as discussed above if the PIK does not match to the Census Numident then the PIK is most likely not eligible to work and may also be a part of the "AA" group. Second, for workers that match to the Census Numident we acquire information on the date of birth (DOB), date of first contact with SSA (DOC), and reported date of death (DOD). These three pieces of information are used to identify PIKs where the original SSN holder is unlikely to be the employee. The DOB criterion identifies PIKs that have jobs active before the age of 14 or after the age of 82. The DOC criterion identifies PIKs that have jobs active more than 1 year prior to the SSN being issued by SSA. The DOD criterion identifies jobs active more than 1 year after the date of death. These three pieces of information are

combined into a single "age_ok" variable. A value of "111" identifies PIKs that pass all three criteria, while any other value results in the PIK being selected for further processing.

Table 2: Pseudo-PIK Process Record Candidate Groups

Row	Record Type	Record Candidate Group	Age	Jobs	Not on CNUM	Not on CNUM (AA)	Workers (PIK)	Share PIK
1	0000	Not Used	0	0	0	0	XXX,XXX,XXX	XX.XX
2	0010	Not on Numident	0	0	1	0	XX,XXX,XXX	X.XX
3	0011	AA-Pseudo-PIK	0	0	1	1	XXX,XXX	X.XX
4	1000	Age-Pseudo-PIK	1	0	0	0	XX,XXX,XXX	X.XX
5	0110	Jobs-Pseudo-PIK	0	1	1	0	XX,XXX	X.XX
6	0100	Jobs-Pseudo-PIK	0	1	0	0	XXX,XXX	X.XX
7	1100	A&J-Pseudo-PIK	1	1	0	0	XXX,XXX	X.XX
Total							XXX,XXX,XXX	

The age, jobs, Census Numident match status, and the presence of an "AA" in the identifier are used to classify each PIK into one of seven record candidate sub-populations as shown in Table 2

The rows of Table 2 show that the pre-selection process is highly successful with about XX% of the PIKs excluded from further analysis (rows 1-2). Only four of the sub-populations (rows 4-7),XX% of the records, pass through the pseudo-PIK "active" selection modules since all PIKs in the "AA" group (row 3) automatically receive a pseudo-PIK. The bulk of the remaining record candidate PIKs are in the age group (row 4, XX% of the records that pass through the "active" Pseudo-PIK selection modules) with the remainder in the jobs group.

Pseudo-PIK Process Overview

The "too many jobs" and "age_ok" modules form the core of the Pseudo-PIK record selection algorithm. PIKs that match to the Census Numident are subject to both the "too many jobs" and the date-of birth/date-of-first contact with SSA tests ("age_ok"). The PIKs that do not match have no reported demographic characteristics and are subject only to the "too many jobs" test. We discuss the "too many jobs" module first. We realized early on that the probability of being a "too many jobs" PIK varies dramatically by state and industry, although the state heterogeneity largely reflects differences in the types of industries active in a state. For example, California has a large movie and TV industry, while Delaware has a large industry supporting companies that incorporate in the state. Although some industries are more prevalent in certain states, we assume that workers in each industry are similar (in terms of the criteria needed to identify a PIK used by multiple persons) across states and focus primarily on industry. This isn't meant to imply we completely disregard geographic differences, but geography is not a direct part of the PIK selection process.

Without a readily available counterfactual, we first identify "suspect" industries where the use of PIKs by multiple persons appear to be the most prevalent by using information on the number of

jobs in a quarter and the "earnings ratio". To calculate the number of jobs criteria, we first select job quarters associated with PIKs who have more than 12 jobs, placing each job quarter into one of three bins; Bin 3: 13-20, Bin 4: 21-50, and Bin 5: 51 or more jobs in the quarter. For each bin, we calculate the share of job quarters in each NAICS 4-digit industry[6]. To be selected, an industry must have a higher share of job quarters in each bin relative to the share of job quarters in a separate reference bin (job quarters associated with PIKs that have 1-6 jobs in the quarter). This metric allows us to identify industries that have relatively large shares of job quarters in bins 3-5 (compared with both other industries and workers in the same industry with 1-6 jobs in the quarter). Certain industries such as "7225: Restaurants and Other Eating Places" and "5617: Services to Buildings and Dwellings" have relatively high shares of job quarters associated with PIKs with many jobs.

However, using information on the number of jobs per quarter is not enough. For example, to avoid selecting industries such as "5511: Management of Companies and Enterprises" where workers legitimately have many jobs (especially in Delaware) we also use an earnings ratio (eratio) measure that compares actual quarterly earnings for PIKs with many jobs (more than 12) with the expected quarterly earnings for those jobs if each job was held by a different worker who themselves had 6 or less jobs (representing the prototypical PIK selected for the pseudo-PIK process). The eratio metric relies on the finite number of hours a worker has available in a quarter. As the number of jobs increases, a single worker will in most cases work fewer hours in each additional job. However, if a PIK is shared by multiple workers the same constraint does not apply. For example, a single worker with 15 jobs should have total earnings in a quarter substantially less than the total earnings for 15 different workers with15 observationally similar jobs.

We estimate the eratio (equation 1) by first calculating median real quarterly earnings for job quarters where the PIK has 6 jobs or less by job type s. The XXX,XXX job types s are created by the full interaction of the following set of job characteristics: job active pattern (single, continuous, full quarter), state, 4-digit NAICS, number of establishments, and firm size class. We use the resulting median earnings table to estimate the eratio for each PIK quarter selected for "too many jobs" processing. We calculate the denominator of the eratio by looking up the median quarterly earnings for each job j in quarter q by its characteristic type s (using the $type$ function) and summing across all jobs in the quarter J_{iq}. The numerator is simply the sum of real actual quarterly earnings across all jobs for PIK i in quarter q.

$$eratio_{iq} = \frac{\sum_{j}^{J_{iq}} earn_{iqj}}{\sum_{j}^{J_{iq}} medearn_{s=type(iqj)}} \quad (1)$$

Figure 4 shows the median estimated eratio by the number of jobs per quarter. The eratio is approximately 1 when a PIK has a single employer in the quarter and falls linearly up to about 4 jobs per quarter. This is the expected behavior if a PIK is used by a single person. However, as the

[6] The 4-digit industries were identified using 2017 NAICS codes, as provided on the ECF. ICF processing will use the 2017 codes for this purpose going forward. The industry lists may be updated to later NAICS coding systems in the future, but it is not required for regular production.

number of jobs per quarter increases the decrease in the eratio slows and by 9 jobs per quarter begins to increase. By 20 jobs per quarter the eratio is stable at about X.X, suggesting that the workers associated with PIKs with many jobs tend to have lower earnings than PIKs with similar types of jobs *not* associated with many jobs. The results in Figure 1 also suggest that PIKs with 10, 20, or more jobs in a quarter are noticeably more likely to have multiple persons using the same PIK.

We assume PIKs with many jobs and the distribution of eratio values provide useful information about the within industry incidence of multiple persons using the same PIK. To better differentiate across industries, we focus on the 25th percentile of the within industry eratio distribution. The 25th percentile has more useful variability than higher percentiles since all industries have at least some PIKs used by multiple persons. Putting everything together, Table 3 shows the suspect industry status for the top 5 industries by size as well as 3 other selected industries.

Insert Figure 4: Median Eratio by the Number of Jobs per Quarter (pending DRB approval)

Table 3: Selected 2017 NAICS 4 Digit Suspect Industry Classifications

					Pct Change in Bin X Industry Share - Bin 1 Industry Share			
NAICS 4 (2017)	Label	Size Rank	Suspect Industry	Share in Bins 3-5	Bin 3	Bin 4	Bin 5	Mean eratio Bins 3-5
7225	Restaurants	XX	1	X.XX	X.XX	X.XX	X.XX	X.XX
6111	Elementary and Secondary Schools	XX	0	X.XX	-X.XX	-X.XX	-X.XX	X.XX
5613	Employment Services	XX	1	X.XX	X.XX	X.XX	X.XX	X.XX
6221	General Medical Hospitals	XX	0	X.XX	-X.XX	-X.XX	-X.XX	
6113	Colleges and Universities	XX	0	X.XX	-X.XX	-X.XX	-X.XX	X.XX
5511	Management of Companies	XX	0	X.XX	-X.XX	X.XX	X.XX	X.XX
2361	Residential Construction	XX	1	X.XX	X.XX	X.XX	X.XX	X.XX
2362	Nonresidential Construction	XX	0	X.XX	X.XX	X.XX	X.XX	X.XX

Two out of the top 5 industries are labeled as suspect; 7225 "Restaurants" and 5613 "Employment Services." The share of job quarters associated with PIKs who have more than 12 jobs is X.XX for industry 7225 and X.XX for industry 5613. In addition, both industries 7225 and 5613 have larger shares of workers with 12 or more employees relative to other industries compared with their relative shares of workers with 6 or more employees. The eratio for both industries is also relatively

high with 7225 having a mean eratio of X.XX and 5613 having a mean eratio of X.XX.[7] Industry 5511 "Management of Companies" provides an interesting comparison to both industries 7225 and 5613. Industry 5511 has a large share of job quarters associated with PIKs who have more than 12 employers (X.XX), but not all of the bins have a relatively high industry share (the Bin 3 value is −X.XX). The eratio is also very low (X.XX) for job quarters associated with PIKs that have 12 or more jobs, suggesting these are mostly very low earning short duration jobs. The last comparison is between residential (2361) and nonresidential (2362) building construction. These industries have somewhat similar characteristics although the eratio values are below the cutoff for nonresidential construction, resulting in a not suspect label. The differences in unionization, owner base, and skillsets required in the nonresidential sector suggest the not-suspect classification is warranted.

Insert Figure 5: Share of Jobs in Suspect Industries by Age at the Beginning of Each Job (pending DRB approval)

The suspect industry designation appears to be informative based on inspection of the full list of industry classifications in Appendix A, but objective criteria also support the suspect designations. An interesting pattern emerges if we assign all jobs where the PIK matches to the Census Numident an age at the beginning of each job; both young and old workers are employed primarily in suspect industries. In Figure 5 we can see that from age 25 to 65 the share of jobs in suspect industries varies between X.XX to X.XX, however as we move further away from the 25 to 65 age range the share of workers in suspect industries increases dramatically. The results are especially puzzling for older workers as many of the jobs in suspect industries require a high level of physical activity as reflected in the downward trend in the share in of jobs in suspect industries from ages 35 to 65. The most likely explanation is that the large increase in the share of jobs in suspect industries for young and old workers is primarily due to these workers using a PIK originally issued to someone else.

Figure 6: Not-Suspect Industries Requiring Special Handling					
NAICS 4 (2017)	Label	Jobs High	Jobs Very High	Age	Farm
11XX	Agriculture, Forestry, Fishing, and Hunting				X
5121	Movie and Video Production	X		X	
5122	Sound Recording	X		X	
5151	Radio and Television Broadcasting	X		X	
5239	Other Financial Investment Activities	X	X		
5259	Other Investment Pools and Funds	X	X		
5311	Lessors of Real Estate	X			

[7] The rule for an industry to be labeled suspect is that the percent difference between bins 3, 4, and 5 industry shares compared with the Bin 1 industry share must all be positive and the eratio for bin 3 must be greater than 0.4 and the eratios for bins 4 and 5 must be greater than 0.6.

5331	Lessors of Nonfinancial Intangible Assets	X	X		
5412	Accounting, Bookkeeping, and Payroll	X			
5417	Research and Development			X	
5418	Advertising and Public Relations	X			
5511	Management of Companies	X	X		
6212	Offices of Dentists	X	X		
7111	Performing Arts Companies	X		X	
7113	Promoters of Performing Arts and Sports	X		X	
7114	Agents and Managers	X		X	
7115	Independent Artists, Writers, and Performers	X		X	

In Figure 6, we list the not-suspect industries that require special handling in the pseudo-PIK job detection phase. The "Jobs High" column shows the industries where workers tend to have many (legitimate) employers. The "Jobs Very High" column identifies industries where workers have even higher numbers of employers relative to the industries that are only in the "Jobs High" column. The "Age" column identifies industries where workers frequently receive payments tracked by the UI system out of the normal age ranges, often due to royalty, commission, or residual payments commonly found in the entertainment industry.[8] The "Farm" column is used to allow for legal agricultural jobs for age 12-13 years old workers.

Both the "too many jobs" and the "age_ok" modules use industry information to better tailor the selection algorithm to each job. For example, jobs in suspect industries receive lower pseudo-PIK selection parameter values given the relatively high probability that a PIK with many jobs is used by multiple persons. In industries where high levels of multiple job holding are common, we use higher cutoff values to reduce the likelihood of selecting PIKs with many jobs held by the same worker.

For PIKs that meet the "too many jobs" pre-selection criteria, in the selection phase we attempt to identify quarters in their work history where the PIK is used by multiple workers. The at-risk period begins the first quarter when the PIK has 7 or more employers and ends in the last quarter where the worker has 7 or more employers, quarters outside this period are not used in the selection process. During the at-risk period, each PIK quarter receives a score based on a principal components analysis. The principal components are estimated using all at-risk quarters for all pre-selected "too many jobs" PIKs with less than 800 jobs (FID or SEIN) and are based on the following quarterly variables: number of jobs, number of full-quarter jobs, number of consecutive quarter jobs, number of new jobs, and the sum of earnings for the 3rd to the J_{iq} lowest earnings jobs.[9] The

[8] Residual payments for actors and other entertainers are tracked by the UI system in many states. These payments are often made non-contemporaneously and are inheritable, creating the appearance of active labor force participation. Unlike other industries, workers in the entertainment industry also often have legitimate employment opportunities at very young (child actors) or old ages. Scientific research entities occasionally employ young or old workers in research studies (one example, is employing under age 14 workers to detassel corn in Nebraska).

[9] If a PIK is used by multiple persons, one of the persons was likely to have been originally assigned the SSN. We assume that the two highest earning jobs are most likely to be held by the person assigned the SSN.

first principal component is a single index that captures as much of the joint variability as possible across the characteristics making it easy to use a single cutoff value. In our case, all the index weights for the first principal component are positive, implying the higher the index value, the higher the probability the PIK is used by multiple persons.

The pseudo-PIK detection algorithm uses two passes through the PIK quarter data. For the first pass, only PIKs that have relatively few job quarters (<25) in "Jobs Very High" industries are eligible. Within this group, two different selection algorithms are used, one for PIK quarters that have at least one job in a "Jobs High" industry (indx=1) and those that do not (indx=0). For PIK quarters with indx=0, the share of jobs in suspect industries must be greater than 0.25 and the first principal component index must be greater than 1.4 for the quarter to be selected. For PIK quarters with indx=1 the first principal component must be greater than 6 (suspect industry information is not used). There is also an additional rule for this group that reverses the selection if the largest number of jobs in the quarter are in a "Jobs Very High" industry and the number of jobs in the "Jobs High" industries is greater than 5. The additional rule limits PIK quarter selections for the group that has some, but not a large amount of exposure to "Jobs High" industries.

Before the next pass the sum of quarterly selections from the first pass is calculated for each PIK. This total represents the number of quarters the PIK is likely to have been used by multiple persons. The second pass refines the selections from the first pass. As in the first pass, PIKs that have relatively few job quarters (<25) in "Jobs Very High" industries are handled separately. If the number of jobs for the PIK is less than 600 and the highest first principal component is less than 14 the PIK is eligible to have the quarter selection reversed, if at least one of the values is higher than those limits the quarter is selected and no further second pass processing is done. If indx=0 and the number of quarters the PIK is likely to have been used by multiple persons is less than 3 and the number of jobs in the quarter is less than 50 than the selection is reversed. If indx=1 and the number of quarters the PIK is likely to have been used by multiple persons is less than 3 and the number of jobs in the quarter is less than 100 than the selection is reversed. For PIKs with 25 or more job quarters in "Jobs Very High" industries and more than 800 jobs (overall) the PIK quarter is selected. If after the first and second passes at least 3 quarters are selected by the "too many jobs" module, all jobs for the PIK are selected to receive a Psuedo-PIK.[10]

The "age_ok" job selection module runs separately from the "too many jobs" module and uses the date of birth (DOB), date of first contact with SSA (DOC), and the reported date of death (DOD) from the Census Numident to identify jobs active during time periods when the PIK is unlikely to participate in the labor market. To be a candidate, the PIK must match to the Census Numident and

---

However, even if this assumption is false, the most important earnings-based indicator of multiple persons using a PIK is relatively high earnings in lower earning jobs.

[10] We attempted to identify time periods within a work history where multiple persons are using the same PIK, but it proved difficult to identify reliable boundaries. Given that the number of PIKs meeting the "too many jobs" criteria is relatively small, most persons hold only 1 or 2 jobs in a quarter, and the number of jobs for the selected PIKs is typically large (minimizing the number of jobs likely to be incorrectly selected for pseudo-PIKs relative to those correctly identified) we chose to include all jobs.

meet the "age_ok" pre-selection criteria.[11] The "age_ok" job selection module uses the suspect industry table as well as the "Age" and "Farm" industry selections for jobs in not-suspect industries as shown in Figure 6 to alter the age-based selection rules. The DOB criteria identify jobs active before the age of 14 or after the age of 99, however unlike the pre-selection "age_ok" criteria industry-based exceptions are made. If the job is in an agriculture industry, then jobs active when the PIK is 12 or 13 are not selected. If the job is active in a suspect industry after the age of 82 then the job is selected.

Insert Figure 7: UI and ACS Annual Employment by Age (pending DRB Approval)

The lower age cutoff for jobs in suspect industries is driven by two results. Looking back at Figure 5, we see that around XX% of job quarters for workers 85 or older are in suspect industries, almost 3 times higher than for prime-age workers. The age of 85 also roughly corresponds with the age where the gap between the ACS and LEHD annual employment counts significantly diverge. The divergence is clearly seen in Figure 7 by comparing the "LEHD PIKs" line with the "ACS Workers" line. The "LEHD age_ok Adjusted" line shows annual employment after removing jobs in suspect industries active after the age of 82. Removing jobs in suspect industries largely corrects the divergence between the ACS and LEHD employment counts for older workers.

The DOC criteria select PIKs that have jobs active more than 1 year prior to the SSN being issued by SSA. The DOC criteria is designed to select jobs active before the PIK is eligible to work. The DOD criteria identify jobs active more than 1 year after the date of death. The final age-based adjustment is for jobs in industries listed in the "Age" column in Figure 6. For jobs in "Age" industries, PIKs with jobs active between the ages of 0-13 and jobs for PIKs reported dead by SSA are not selected. This is done primarily to create an exception for workers in the entertainment industry who have higher young and old worker labor force participation than other groups.

Results

The results in this section use PIK earnings history information from the LEHD 2022 snapshot, which includes data from 1985 quarter 4 to 2023 quarter 1. Some of the figures and tables with annual statistics use information only from 2000 to 2022 due to the time-series impact of state entry prior to 2000. Table 4 contains our first results, which shows tabulations for the PIKs pre-selected for Pseudo-PIK processing compared with the set of final PIKs selected. The rows of Table 4 identify the record type, with the first panel showing counts and the second panel showing shares. The "PIK Pre-Select" column replicates information from the "Workers (PIK)" column in Table 2 and is the baseline for the rest of Table 4. In the Pseudo-PIK process program flow, the pre-select module runs first and selects PIKs that are candidates for either the "too many jobs" and/or the "age_ok" modules. About XX% of the PIKs (rows 1 and 2) are not chosen by the pre-select

[11] The "age_ok" pre-selection module examines all jobs for a PIK, while the "age_ok" job selection module applies the same criteria to each job separately. The rules and logic are identical except for the industry-based exceptions described here.

module, with the records in row 3 all receiving a pseudo-PIK. Ideally, after the pre-select module runs all of the PIKs would be in row 1 where the PIK matches to a record on the Census Numident, has reasonable levels of job holding every quarter, all jobs are active during the expected age range, all jobs start less than 1 year prior to the SSN having been issued by SSA, and all jobs end no more than 1 year after the reported date of death. The PIKs in rows 4-7 of the "PIK Pre-Select" column fail at least one of these conditions and are the primary focus of the Pseudo-PIK selection modules ("too many jobs" and "age_ok").

The "PIK SEIN Pre-Select" column shows the job counts associated with different types of pre-selected PIKs.[12] Overall, the PIK SEINs in row 1 (Count) tend to have a larger share of jobs than the pre-selected PIKs in row 1, but the increase is relatively small. As discussed earlier, the PIKs in row 2 "Not on Numident" have very few jobs, relative to the PIKs in other rows (in Figure 3A, we show that about XX% have only 1 job). PIKs in rows 3, 5, 6, and 7 all have a large increase in shares in the "PIK SEIN Pre-Select" column relative to the "PIK Pre-Select" column, showing that the average number of jobs for the PIKs in these rows is relatively high.

[12] Although the pseudo-PIK process uses the PIK FID as the job identifier, PIK FIDs are ephemeral and consist of sets of SEIN's that persist across production runs. Ultimately, although the pseudo-PIK process uses PIK FIDs as the job identifier, the final output is a list of PIK SEINs associated with the identified PIK FIDs. Occasionally the set of SEINs associated with a FID can change, resulting in a change in the PIK SEIN list even though the selected PIK FID itself has not changed. The results presented here use the SEIN to identify a job, which inflates the job counts relative to using the PIK FID, however using the SEIN makes the analysis more straightforward without significantly affecting the results.

Table 4: Selection Counts and Rates by Record Type									
Count									
Row	Record Type	Record Category	PIK Pre-Select	PIK SEIN Pre-Select	PIK Select	PIK SEIN Select	PIK SEIN Year Select	PIK Select Rate	PIK SEIN Select Rate
1	0000	Not Used	XXX,XXX,XXX	X,XXX,XXX,XXX	XXX,XXX,XXX	X,XXX,XXX,XXX	X,XXX,XXX,XXX	X.XXX	X.XXX
2	0010	Not on Numident	XX,XXX,XXX	XX,XXX,XXX	XX,XXX,XXX	XX,XXX,XXX	XX,XXX,XXX	X.XXX	X.XXX
3	0011	AA-Pseudo-PIK	XXX,XXX	XX,XXX,XXX	XXX,XXX	XX,XXX,XXX	XX,XXX,XXX	X.XXX	X.XXX
4	1000	Age-Pseudo-PIK	XX,XXX,XXX	XXX,XXX,XXX	XX,XXX,XXX	XX,XXX,XXX	XXX,XXX,XXX	X.XXX	X.XXX
5	0110	Jobs-Pseudo-PIK	XX,XXX	X,XXX,XXX	X,XXX	XXX,XXX	X,XXX,XXX	X.XXX	X.XXX
6	0100	Jobs-Pseudo-PIK	XXX,XXX	XX,XXX,XXX	XX,XXX	X,XXX,XXX	X,XXX,XXX	X.XXX	X.XXX
7	1100	A&J-Pseudo-PIK	XXX,XXX	XX,XXX,XXX	XX,XXX	X,XXX,XXX	X,XXX,XXX	X.XXX	X.XXX
Total			XXX,XXX,XXX	X,XXX,XXX,XXX	XXX,XXX,XXX	X,XXX,XXX,XXX	X,XXX,XXX,XXX		
Share									
Row	rtype	Record Category	PIK Pre-Select	PIK SEIN Pre-Select	PIK Select	PIK SEIN Select	PIK SEIN Year Select	PIK Select Rate	PIK SEIN Select Rate
1	0000	Not Used	XX.XXX	XX.XXX	XX.XXX	XX.XXX	XX.XXX	X.XXX	X.XXX
2	0010	Not on Numident	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX
3	0011	AA-Pseudo-PIK	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX
4	1000	Age-Pseudo-PIK	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX
5	0110	Jobs-Pseudo-PIK	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX
6	0100	Jobs-Pseudo-PIK	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX
7	1100	A&J-Pseudo-PIK	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX	X.XXX

Continuing to the right in Table 4, we have the "PIK Select" column. This column shows the final count of PIKs selected by the Pseudo-PIK selection modules. The related "PIK Select Rate" column shows the average rate of pre-selected PIKs that will receive a pseudo-PIK. For example, approximately X.X% additional PIKs in row 1 are not assigned a pseudo-PIK compared with the PIKs pre-selected. Almost all the PIKs in row 4 "Age-Pseudo-PIK" are selected (X.XXX), but the pseudo-PIK PIK selection rate for the "too many jobs" module is much lower for rows 5,6, and 7 with rates of X.XXX, X.XXX, and X.XXX respectively. For example, the number of PIKs in row 6 pre-selected for the "too many jobs" module is XXX,XXX while only XX,XXX PIKs are ultimately assigned a pseudo-PIK, a very low rate. The "PIK SEIN Select" and the "PIK SEIN Select Rate" columns provide a different perspective, weighting the result by the number of jobs per PIK. When focusing on jobs (PIK SEINs), the selection rates are much higher for the "too many jobs" module, primarily due to the relatively large number of jobs associated with each of the PIKs. For example, the selection rate for row 6 PIKs is X.XXX, but the selection rate for PIK SEINs is X.XXX. When a PIK is selected by the "too many jobs" module all the jobs are selected and each PIK has many jobs, generating a significantly higher PIK SEIN selection rate. For the "age_ok" module typically only a small subset of the jobs for each PIK are selected, reducing the PIK SEIN selection rate relative to the PIK selection rate (X.XXX for PIKs compared with X.XXX for PIK SEINs). Overall, the pseudo-PIK active selection modules (represented by rows 4-7) choose at least one job for X.X% of PIKs, select X.X% of all PIK SEIN records, and select X.X% of all PIK SEIN Year records.

Figures 8A and 8B show the counts and shares of pseudo-PIK jobs over time from 2000 to 2022. In Figure 8A we show the total count of records that do not match to the Numident (including AA records) as well as the jobs (PIK SEIN) that actively receive a pseudo-PIK by year. The total for both groups is about X.X million jobs in 2000 rising to a peak of about X.X million jobs in 2006 just before the start of the Great Recession. The total then declines noticeably during and shortly after the Great Recession staying roughly constant at around X.X million from 2010 to 2020 before starting to increase in 2021. The two components of the total have very different trends. The "Active Pseudo-PIK" job counts are generally declining over the period while the "Not on Numident (incl AA)" job counts are increasing with the latter group exceeding the former for the first time in 2021. It isn't clear why the share of each group is changing over the period, but changes in the mix of states using E-Verify and changes in the locations of illegal immigrant inflows are possible explanations.

Figure 8B shows the two groups from Figure 8A in more detail, plotting the component PIK record types (rows 2-7 in Table 4) over time. For the "Not on Numident (incl AA)" group from Figure 8A, both component shares increase substantially over time, with the largest relative increase coming from the "AA-Pseudo-PIK" group. For the "Active Pseudo-PIK" group almost all the decrease comes from a large decline in the share of PIK SEIN's receiving a pseudo-PIK from the "age_ok" module. The share of "Age-Pseudo-PIK" PIK SEINs declines from X.XX in 2000 to X.XX in 2022 with the rest of the component shares of the "Active Pseudo-PIK" group relatively stable over the period.

Insert Figure 8A: Count of Pseudo-PIK Process Active PIK SEINs by Year (pending DRB approval)

Insert Figure 8B: Share of Pseudo-PIK process Active PIK SEINs by Year and Record Type (pending DRB approval)

Figure 9 shows the share of pseudo-PIK jobs (PIK SEIN) in a state using the same groups and sample (2000-2022) as in Figure 8A. We see substantial variation in the share of pseudo-PIK jobs around the first bar in Figure 9 which displays the share for the U.S. The overall share of pseudo-PIK jobs is X.XXXX with the "Not on Numident (incl AA)" share equal to X.XXXXX and the "Active Pseudo-PIK" share equal to X.XXXXX. The five states with the largest shares of pseudo-PIK jobs are NC, CA, WA, NV, and IL. Three of the states are in the West, one is in the South, and one is in the Midwest. North Carolina (NC) has an extremely large share of Pseudo-PIK jobs (including records that do not match to the Numident), X.XXX, over XX% larger than California (CA), the state with the next largest share. Even more puzzling is that almost all the gap is due to NC having an extremely large share of AA records. The five states with the smallest share of pseudo-PIKs are PR, WV, ME, HI, and VT. Two of the states are islands surrounded by water, two are in New England, and the final state is in the Mid-Atlantic region. Only 9 of the 52 states have a share of Pseudo-PIKs greater than the national average with 5 states in the West (CA, NV, UT, OR, WA), 2 along the border with Mexico (CA, TX) and the other 3 (IL, NC, NJ) scattered across the rest of the U.S. Overall, states along the West Coast and in the Southwest tend to have the highest share of Pseudo-PIK jobs, but there are noticeable exceptions such as IL, NC, and NJ.

Figure 9 also includes information on the Ratio of "Active Pseudo-PIK" jobs to "Not on Numident (incl AA)" jobs. The two most extreme outliers are NC and AZ with NC having a relatively low ratio (a large share of "Not on Numident (incl AA)" jobs) and AZ having the opposite (a large share of pseudo-PIK jobs actively assigned by the Pseudo-PIK selection modules). Unfortunately, it isn't clear what is driving the results for either state. The states with the lowest shares of "pseudo-PIK" jobs tend to have somewhat higher ratios, but there is substantial variability across states. In most of the states with the highest ratios the share of both types of records, "Not on Numident (incl AA)" and "Active Pseudo-PIK", are relatively high. However, as the overall share of pseudo-PIKs in a state declines the share of "Not on Numident (incl AA)" jobs declines even faster, increasing the ratio.

Insert Figure 9: Share Active Pseudo-PIK Process PIK SEIN Year Records of Total by State and Type ; Ratio Active Pseudo PIK/Not on Numident

Although the overall job selection rate for the Pseudo-PIK process is low, we showed in Figure 9 how the share of pseudo-PIK jobs varies substantially by geography (state). In the next three tables, we explore differences in the share of pseudo-PIK jobs by industry. In Table 5, we examine the "actively" selected pseudo-PIKs in detail for selected industries.[13] The rows of Table 5 are composed of the top 5 PIK SEIN pre-selected industries, the industries in the "Jobs High" column in Figure 6, and a few other comparison industries (6113 and 9221). NAICS 2017 industry 7225 "Restaurants" has the largest number of jobs (PIK SEIN) pre-selected for the pseudo-PIK process (see the "Pre-Select Rank" column). Industry 7225 also has the largest number of jobs selected to receive a pseudo-PIK; XX,XXX,XXX jobs will receive a pseudo-PIK vs XX,XXX,XXX jobs pre-selected, a decrease of XX,XXX,XXX jobs. The pre-select module selects XX.X% of the jobs in industry 7225 for pseudo-PIK processing, but only X.X% of the jobs in the industry were ultimately selected to receive a pseudo-PIK. Although industry 7225 is a suspect industry, the final selection rate is not particularly high compared with, for example, industry 1151 "Farming (Crop Support)", the industry with the third largest number of pre-selected jobs. Industry 1151 has XX.X% of jobs pre-selected for the pseudo-PIK processing with XX% of the jobs in industry 1151 ultimately selected. The two industries provide a nice comparison, the "Restaurants" industry has a large number of selections, but this is largely due to its size, the selection rate is not particularly high compared with the national average of X.X%, while the "Farming (Crop Support)" industry has XX% of jobs selected to receive a pseudo-PIK. [14]

As described previously in the section discussing the pseudo-PIK algorithm, jobs in suspect industries are more likely to be selected by the pseudo-PIK process than jobs in industries where PIKs tend to legally have many employers. For example, the results for industry 5121 "Film and Video" show the impact of altering the pseudo-PIK selection algorithm for specific industries with high rates of legitimate multiple job holding. Industry 5121 has the 9th largest number of pre-selected PIK SEINs, however only the 127th largest number of PIK SEINs selected by the pseudo-PIK process. A person with a substantial share of jobs in the "Jobs High" category in a quarter is expected by the Pseudo-PIK algorithm to have a relatively large number of jobs, reducing the likelihood a PIK quarter is selected. The impact of the algorithm changes results in a very small PIK selection rate for industry 5121 of X.XXX.[15]

---

[13] In Table 5 we are attempting to show how the pseudo-PIK selection modules deal with jobs in different industries. For this table only, we do not include the "Not on Numident (incl AA)" group. The pseudo-PIK algorithm itself does not "actively" affect the assignment of pseudo-PIKs within the two not on Numident sub-groups; all PIKs in the "Not on Numident" group are not assigned pseudo-PIKs while all PIKs in the AA group are assigned pseudo-PIKs.

[14] Industries and areas with a large share of pseudo-PIKs are at risk to have significant changes to publicly released LEHD statistics such as the QWI, but the ultimate impact will depend on the imputed PIK characteristics.

[15] People who work in entertainment have a low level of PIKs that do not match to the Numident, suggesting that the level of PIKs associated with multiple persons is low. Persons who work in entertainment often also receive legitimate payments at ages when the typical worker is not in the labor force. Child actors, residuals, royalties, and commissions result in spurious pseudo-PIK job selections with the normal "age_ok" parameters. For PIKs in industry 5121 we relax the age bounds (0-99 is allowed) and adjust the parameters in the "too many jobs" module to allow for more jobs in a quarter.

Table 5: Selected NAICS 4 Digit Ranks, PIK SEIN Counts, and Within Industry Shares (Pre and Final Selection)											
NAICS 4 (2017)	Label	Pre-Select Rank	Select Rank	Rank Diff	Suspect	Jobs High	Pre-Select Count	Select Count	Count Diff	Share Pre-Select	Share Select
7225	Restaurants	X	X	X	1	0	XX,XXX,XXX	XX,XXX,XXX	-XX,XXX,XXX	X.XXX	X.XXX
5613	Employment Services	X	X	X	1	0	XX,XXX,XXX	X,XXX,XXX	-XX,XXX,XXX	X.XXX	X.XXX
1151	Farming (Crop Support)	X	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
5617	Services to Buildings	X	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
1113	Farming (Tree and Nut)	X	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
2381	Construction (Foundation)	X	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
5121	Movie and Video Production	X	XXX	-XXX	0	1	X,XXX,XXX	XX,XXX	-X,XXX,XXX	X.XXX	X.XXX
5412	Accounting	XX	XX	-X	0	1	X,XXX,XXX	XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
2383	Construction (Finishing)	XX	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
2361	Construction (Residential)	XX	X	X	1	0	X,XXX,XXX	X,XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
6113	Colleges and Universities	XX	XX	-XX	0	0	X,XXX,XXX	XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
1112	Farming (Vegetable and Melon)	XX	XX	XX	1	0	X,XXX,XXX	XXX,XXX	-XXX,XXX	X.XXX	X.XXX
1119	Farming (Other)	XX	XX	XX	1	0	X,XXX,XXX	XXX,XXX	-XXX,XXX	X.XXX	X.XXX
6212	Dentists	XX	XX	-XX	0	1	X,XXX,XXX	XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
5511	Management of Companies	XX	XX	-XX	0	1	X,XXX,XXX	XXX,XXX	-X,XXX,XXX	X.XXX	X.XXX
5311	Lessors (Real Estate)	XX	XX	XX	0	1	XXX,XXX	XXX,XXX	-XXX,XXX	X.XXX	X.XXX
7111	Performing Arts Companies	XX	XXX	-XXX	0	1	XXX,XXX	XX,XXX	-XXX,XXX	X.XXX	X.XXX
3152	Cut and Sew Apparel	XX	XX	XX	1	0	XXX,XXX	XXX,XXX	-XXX,XXX	X.XXX	X.XXX
5418	Advertising and Public Relations	XX	XX	-X	0	1	XXX,XXX	XXX,XXX	-XXX,XXX	X.XXX	X.XXX
9221	Justice and Safety	XX	XXX	-XX	0	0	XXX,XXX	XX,XXX	-XXX,XXX	X.XXX	X.XXX
5417	Research and Development	XXX	XXX	-XXX	0	0	XXX,XXX	XX,XXX	-XXX,XXX	X.XXX	X.XXX
7113	Promoters (Performing Arts)	XXX	XXX	-XXX	0	1	XXX,XXX	XX,XXX	-XXX,XXX	X.XXX	X.XXX
5239	Financial Investment (Other)	XXX	XXX	-X	0	1	XXX,XXX	XX,XXX	-XXX,XXX	X.XXX	X.XXX
7115	Artists and Writers	XXX	XXX	-XXX	0	1	XXX,XXX	X,XXX	-XXX,XXX	X.XXX	X.XXX
5151	Radio and TV Broadcasting	XXX	XXX	-XXX	0	1	XXX,XXX	X,XXX	-XXX,XXX	X.XXX	X.XXX

We can see similar large impacts of the algorithm changes for other industries in the "Jobs High" category, especially for industries 7111, 5417, 7113, 7115, and 5151. Industries not part of the "Jobs High" category but which are also unlikely to have many PIKs used by multiple persons have similar reductions using the standard algorithm; see for example industries 6113 "Colleges and Universities" and 9221 "Justice and Safety".

In Table 5, we showed results that highlight how the Pseudo-PIK algorithm adapts to different job-holding patterns within industries. For our final set of results, we focus less on the details of the Pseudo-PIK process but instead examine the full set of PIKs that either do not match to the Census Numident or are actively identified by one of the Pseudo-PIK process modules ("too many jobs" or "age_ok"). Our intent here is to give users of LEHD data a better understanding of the industries with the highest potential impact of pseudo-PIKs and pseudo-PIK like records, recognizing that the expected impact from each record type on LEHD statistics is about the same. For example, in the combined group almost every PIK or pseudo-PIK has only one job (XX% of the PIKs have one job in the "Not on Numident" group and XXX% of the pseudo-PIKs in both the "AA-Pseudo-PIK' and the "Active Pseudo-PIK" groups have only one job) with each job receiving demographic characteristic implicates from similar imputation models.

Table 6 shows the XX NAICS 4-digit industries with the largest counts of pseudo-PIK PIK SEIN Year records in descending order.[16] The XX industries in Table 6 contain just over XX% of all pseudo-PIK PIK SEIN Year records.[17] Table 7 is intended to complement Table 6, showing all industries with high shares of pseudo-PIK PIK SEIN Year records, even if the industries are relatively small. When interpreting Table 6 it is useful to keep in mind that the overall pseudo-PIK PIK SEIN Year share is X.XXX. Most of the industries in Table 6 have pseudo-PIK shares over this value, but some do not. For example, industry 4451 "Grocery Stores" has a pseudo-PIK share of X.XXX. This share ranks XXXth out of 311 NAICS 2017 4-digit industries. Grocery stores are common in all states and in XX out of XX the count of pseudo-PIKs for industry 4451 ranks in the top XX, however in none of the states does the share rank in the top XX. These statistics make it clear that industry 4451 is in the top XX largest industries primarily due to its size, not the share of pseudo-PIKs.

Industry 1112 "Farming (Vegetable and Melon)" isn't as large as industry 4451, but it has a relatively high share of pseudo-PIKs (X.XXX). This share of pseudo-PIKs is the third highest across all industries (1113 Farming (Tree and Nut) is the highest with a share of X.XXX). Six states have a relatively large 1112 industry and the pseudo-PIK count makes it into the top XX in those states. However, even if the 1112 industry is small in a state, the pseudo-PIK share is typically in the top XX (XX out of XX states). Table 6 provides a nice overview of the industries with the largest pseudo-PIK counts, but some of the smaller industries with relatively high shares of pseudo-PIKs are missing.

[16] Figure 8, Figure 9, Table 6, and Table 7 use a different sample than Tables 2, 4, and 5. For example, Table 4 and Table 5 contain PIK SEIN counts for all years, while Table 6 contains PIK SEIN Year counts for 2000-2022, changing the industry rankings in some cases. For example, industry 1151 ranks 3rd using PIK SEIN and ranks 4th using PIK SEIN Year.

[17] A pseudo-PIK in the context of Tables 6 and 7 includes both actual pseudo-PIK and pseudo-PIK like records (all record types in Table 2 except record type 1).

Table 6: Thirty Largest Pseudo-PIK Process NAICS 4 Digit Industries (Selected Statistics)						
					States in Top 10	
NAICS 4 (2017)	Label	Count	Share	Share Rank	Count	Share
7225	Restaurants	XX,XXX,XXX	X.XXX	XX	XX	X
5613	Employment Services	XX,XXX,XXX	X.XXX	XX	XX	X
5617	Services to Buildings	X,XXX,XXX	X.XXX	XX	XX	XX
1151	Farming (Crop Support)	X,XXX,XXX	X.XXX	X	X	XX
2381	Construction (Foundation)	X,XXX,XXX	X.XXX	X	XX	XX
1113	Farming (Tree and Nut)	X,XXX,XXX	X.XXX	X	X	XX
2383	Construction (Finishing)	X,XXX,XXX	X.XXX	XX	XX	XX
7211	Traveler Accommodation	X,XXX,XXX	X.XXX	XX	XX	X
2361	Construction (Residential)	X,XXX,XXX	X.XXX	XX	XX	X
4451	Grocery Stores	X,XXX,XXX	X.XXX	XXX	XX	X
2382	Construction (Equipment)	X,XXX,XXX	X.XXX	XX	XX	X
2389	Construction (Specialty Trade)	X,XXX,XXX	X.XXX	XX	XX	X
8111	Automotive Repair	X,XXX,XXX	X.XXX	XX	X	X
1112	Farming (Vegetable and Melon)	X,XXX,XXX	X.XXX	X	X	XX
7139	Amusement Industries	X,XXX,XXX	X.XXX	XX	XX	X
1114	Farming (Greenhouse)	X,XXX,XXX	X.XXX	X	X	XX
2362	Construction (Non-Residential)	X,XXX,XXX	X.XXX	XX	X	X
4244	Grocery Wholesalers	X,XXX,XXX	X.XXX	XX	X	X
8123	Drycleaning and Laundry Services	X,XXX,XXX	X.XXX	XX	X	X
1121	Farming (Cattle)	X,XXX,XXX	X.XXX	X	XX	XX
6111	Elementary and Secondary Schools	XXX,XXX	X.XXX	XXX	XX	X
1119	Farming (Other)	XXX,XXX	X.XXX	X	X	XX
2371	Construction (Utilities)	XXX,XXX	X.XXX	XX	X	X
3118	Manufacturing (Bakery/Tortilla)	XXX,XXX	X.XXX	XX	X	X
7223	Special Food Services	XXX,XXX	X.XXX	XX	X	X
4471	Gasoline Stations	XXX,XXX	X.XXX	XXX	X	X
3116	Animal Slaughtering and Processing	XXX,XXX	X.XXX	XX	X	X
3261	Manufacturing (Plastics)	XXX,XXX	X.XXX	XX	X	X
5412	Accounting	XXX,XXX	X.XXX	XXX	X	X
3219	Manufacturing (Wood)	XXX,XXX	X.XXX	XX	X	X

Table 7: NAICS 4 Digit Industries with more than 10% Share Pseudo-PIKs (Selected Statistics)						
					States in Top 10	
NAICS 4 (2017)	Label	Count	Share Suspect	Share Suspect Rank	Count	Share
5617	Services to Buildings	X,XXX,XXX	X.XXX	XX	XX	XX
1151	Farming (Crop Support)	X,XXX,XXX	X.XXX	X	X	XX
2381	Construction (Foundation)	X,XXX,XXX	X.XXX	X	XX	XX
1113	Farming (Tree and Nut)	X,XXX,XXX	X.XXX	X	X	XX
2383	Construction (Finishing)	X,XXX,XXX	X.XXX	XX	XX	X
1112	Farming (Vegetable and Melon)	X,XXX,XXX	X.XXX	X	X	XX
1114	Farming (Greenhouse)	X,XXX,XXX	X.XXX	X	X	XX
1121	Farming (Cattle)	X,XXX,XXX	X.XXX	X	XX	XX
1119	Farming (Other)	XXX,XXX	X.XXX	X	X	XX
3152	Cut and Sew Apparel	XXX,XXX	X.XXX	XX	X	X
1122	Farming (Hog and Pig)	XXX,XXX	X.XXX	X	X	XX
1123	Farming (Poultry)	XXX,XXX	X.XXX	XX	X	XX
3133	Textile and Fabric Finishing	XXX,XXX	X.XXX	XX	X	X
1152	Farming (Animal Support)	XXX,XXX	X.XXX	XX	X	XX
1153	Farming (Forestry Support)	XXX,XXX	X.XXX	XX	X	X
1129	Farming (Other Animal)	XX,XXX	X.XXX	XX	X	XX
1132	Farming (Forest Nurseries)	XX,XXX	X.XXX	X	X	XX

Table 7 shows the 17 industries with a pseudo-PIK share greater than X.XX. The industries in rows 1-9 are present in Table 6, but the industries in rows 10-17 are new. Although the industries in rows 10-17 are smaller, they are concentrated in two sectors: Agriculture and Manufacturing (clothes and fabric manufacturing in particular). Each of these industries has 3-4 times the share of pseudo-PIKs as the national average. Overall, Tables 6 and 7 show that the industries with the highest share of pseudo-PIKs are clustered in Agriculture and Construction. Many other industries have a large number of pseudo-PIKs, but this is due primarily to the size and not the share of pseudo-PIKs. For example, industry 7225 "Restaurants" has the largest number of pseudo-PIK PIK SEIN Year records with a share of pseudo-PIKs not too far above the national average. However, relative to other NAICS 4-digit industries in Agriculture and Construction the share of pseudo-PIKs is relatively low.

Conclusion

The pseudo-PIK process identifies jobs associated with PIKs used by multiple persons. The pseudo-PIK active selection modules use information from the Census Numident (date of birth, date of first contact with SSA, and date of death), the PIKs job history, and the industries in which those jobs occur to identify PIK FIDs. Each identified job (PIK FID) receives a unique pseudo-PIK that triggers an imputation of all demographic characteristics. The set of actively identified pseudo-PIKs have a closely related sub-population, the set of PIKs that do not match to the Census Numident. Taken together, at least one job for XX.X% of the PIKs on the LEHD EHF will receive imputed demographic characteristics. Given that not all jobs for a PIK are selected, the expected impact on job-based statistics is smaller, with about X.X% of all jobs (PIK SEIN) receiving imputed characteristics due to either a pseudo-PIK or the PIK not matching to the Census Numident.

From a data user perspective, the impact of the pseudo-PIK process should be minimal. After the PIK FIDs are identified, the EHF and ICF are updated to use the new pseudo-PIKs. The ECF is not directly affected, since the firm identifiers have not changed. After the pseudo-PIKs are applied to the EHF, the maximum number of jobs per PIK (both SEIN and FID) is 800. This will make some types of processing easier since the number of jobs per PIK is no longer unbounded.

References

McKinney, Kevin and Stephen Tibbets (2025), Understanding the Longitudinal Employer Household Dynamics (LEHD) Linked-Job Identifier (FID), CES Technical Note