Post-Secondary Employment Outcomes (PSEO)*

Originally Released: July 2, 2019

Last Updated: November 7, 2019

Andrew Foote, Joyce Key Hahn, Stephen Tibbets, and Larry Warren

November 7, 2019

Introduction

Post-Secondary Employment Outcomes (PSEO) are experimental tabulations developed by the Census Bureau in collaboration with post-secondary institutions and state agencies. PSEO data provide earnings and employment outcomes for college and university graduates by degree level, degree major, and post-secondary institution. The current PSEO data are released as a research data product in "beta" form.

The PSEO provide data on earnings and employment for recent graduates of partner colleges and universities. The earnings tabulations were released in March 2018 and are updated as more partners join the program. A second set of tabulations provides industry and location of employment for graduates. These statistics are generated by matching university transcript data with a national database of jobs, using state-of-the-art confidentiality protection mechanisms to protect the underlying data.

Uses

Economic considerations drive a number of college decisions - whether to attend college, where to attend, and what major to select. Given the substantial sums paid by students, especially with loans that must be repaid after leaving school, students want to know whether programs are likely to have a sufficient return to justify their expense. Prospective students would also like to know

*Disclaimer goes here

what labor markets recent graduates are working in and whether or not they are employed in an industry appropriate for their training.

Existing data provide some information on earnings for graduates but have certain limitations. College Scorecard data from the U.S. Department of Education are restricted to federal aid recipients, who may not be representative of the student population. PayScale, a commercial website, publishes earnings by institution and degree, but relies on voluntary self-reported earnings, not generally considered a scientifically valid sampling method. Many states, such as Texas, have matched transcript data to state job records to produce statistics on earnings and employment for graduates. However, state administrative data systems cannot follow students out of state, biasing earnings and employment downward in the matched data.

PSEO statistics also summarize flows of recent college graduates across industries and geographic areas, which is the first time such statistics have been created. These data will be particularly useful for business and state administrations who are interested in how many people are flowing into and out of a state with a given degree level and field.

PSEO statistics are created in a similar way to the state-based matching systems (universal coverage of post-secondary graduate population, longitudinal information on earnings and employment after graduation) but with an important advantage - the ability to follow students across state lines. The PSEO also use cutting edge differential privacy methods to protect the confidentiality of the underlying data, a protection method developed in computer science to bound the privacy risk to individuals from multiple queries to the same database. Differential privacy methods allow the U.S. Census Bureau to release detailed tabulations on student outcomes while minimizing the privacy risk to individuals in the data.

Data Sources

The sample frame for the PSEO is persons who received a degree or certificate from an in-scope institution. Institutions provide Census with a graduation file, which reports the degree type, degree field, graduation date, and institution for any graduating student. Demographic data on students are also provided. The PSEO are created by merging these post-secondary education records with administrative data on jobs collected by the Longitudinal Employer-Household Dynamics

Program (LEHD) at Census. Specifically, we use the Employment History File, and the Employer Characteristics File to generate the PSEO data. Currently, the data cover 2001-2016, but will be updated as more earnings and graduate data become available.

Post-Graduate Population Coverage

Transcript data are provided to Census by higher education systems and individual colleges and universities through data sharing agreements with Census. A list of partner institutions and coverage dates for each system is provided in Appendix Table 2. In the initial pilot phase of the PSEO, only a handful of institutions are represented, but institutional coverage will expand as the program expands.

PSEO tabulations include only graduates of in-scope institutions. Students who enroll but do not graduate are omitted from the statistics. Of these graduates, a very small fraction (less than one percent of graduates) are omitted from the published statistics due to poor quality of the personal identifier. A much larger fraction of graduates is omitted from the earnings and employment outcome statistics because of insufficient labor market attachment in the reference year. For example, a graduate with zero earnings for three quarters of the calendar year but positive earnings in a single quarter will not be included in the earnings statistics or employment counts. These graduates are omitted as the PSEO is intended to reflect earnings and employment for graduates who work throughout the year. More specifics on the labor force attachment restrictions are provided in the earnings section.

Employment Coverage

The LEHD data at Census are a quarterly database of jobs covering over 96% of employment in the United States. The core jobs data are state unemployment insurance (UI) wage records collected via a voluntary federal-state data sharing partnership. These job records are then supplemented with Census surveys and other federal agency administrative records to supply additional information on the characteristics of the workers and firms. This linked employer-employee data for the U.S. are the source data for Census's Quarterly Workforce Indicators (QWI), LEHD Origin-Destination Employment Statistics (LODES), and Job-to-Job Flows (J2J). More information about the LEHD data is available in Abowd et al. (2009).

Private-Industry Employment. Covered private-industry employment in the LEHD data includes most corporate officials, all executives, all supervisory personnel, all professionals, all clerical workers, many farmworkers, all wage earners, all piece workers, and all part-time workers. Workers on paid sick leave, paid holiday, paid vacation, and the like are also covered. Workers on the payroll of more than one firm during the period are counted by each employer that is subject to UI, as long as those workers satisfy the preceding definition of employment. Workers have UI wages filed in every quarter they are covered, even though their wages may not be subject to UI tax in the later months of the year.

Notable exclusions from UI coverage among private sector employers are independent contractors, the unincorporated self-employed, railroad workers covered by the railroad unemployment insurance system, some family employees of family-owned businesses, certain farm workers, students working for universities under certain cooperative programs, salespersons primarily paid on commission, and workers of some non-profits. States have some leeway in designating coverage; for a complete list, see the coverage section of the most recent Comparison of State UI laws.

State and Local Government Employment. Covered employment in the LEHD data includes most employees of state and local governments with the exception of elected officials, members of a legislative body or the judiciary, and some emergency employees.

Federal Government Employment. Federal government workers are not covered by state UI. LEHD uses data from the Office of Personnel Management (OPM) to generate earnings and employment histories for federal workers. The OPM data cover most federal employees but excludes White House officials, members of Congress and the Judiciary, and certain national security agencies, which are excluded for security reasons. Members of the armed forces and the U.S. Postal Service are not covered in OPM data. OPM data currently cover 2000-2015.[1]

UI Coverage across years. Availability of UI data in the LEHD system varies by state. LEHD has data for about ten states in the early 1990s, expanding rapidly to 40 states by the late 1990s, with Massachusetts being the last state to enter the LEHD system in 2010. A continually updated table of state data availability is available here: `https://qwiexplorer.ces.census.gov/loading_`

[1]Releases in 2019 and earlier do not include the full OPM data. For earnings releases, the OPM data cover 2000-2011, while for the flows releases, OPM data is not included. We plan to re-release the data in the future once we integrate earnings from W2 and 1099 files.

status.html.

IPEDS Completions Data

The Graduate Earnings data only release the count of employed graduates in a given cell, but not the overall count of graduates in a cell. To allow users to contexualize the employment count, we supplement our Graduate Earnings tables with data from the Integrated Postsecondary Education Data System, which is maintained by the National Center for Education Statistics in the Department of Education.

IPEDS publishes counts of completions by institution, degree level, degree field, and academic year of graduation.[2] We use these data to approximate the number of graduates in a cell, which we attach to our output data file. Due to the mismatch of timing between academic year and calendar year, the IPEDS counts may not exactly represent the number of graduates in a given cohort. We do not clean the IPEDS data, with the exception of a few expert edits, which we flag in the data; see the flag descriptions in the data schema. There are some cells for which IPEDS does not have any completions data. These are coded as missing, and given the appropriate flag.

Degree, Earnings, and Employment Concepts

Institution, Degree, and Program. Formally, the institution is identified by the 6-digit Office of Post-secondary Education ID (OPEID). A full list of the degree level values is in the following subsection. To classify field of study, we use the Classification of Instructional Program (CIP) codes. For Masters and Doctoral-Research degrees, we classify field of study at the 2-digit CIP level, while for all other degree levels, we classify field of study at the 4-digit CIP level. For each university system, we process the transcript data to standardize variables and update older CIP codes to the most recent classifications (currently 2010 CIP codes). We consider students who earn multiple degrees in the system to be separate observations. Additionally, we consider a student who double-majored as two separate observations, as long as the 6-digit CIP codes are different.[3]

Year Post-Graduation. For all post-secondary graduates, the first year post-graduation is defined

[2] These tables are available on the NCES website at https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx.

[3] Additionally, we do not exclude individuals from a degree if they earn a subsequent degree, but rather consider these degrees to be separate observations in the data.

as the first calendar year following their graduation year. So for a student who graduates in May of 2005, year one begins in January of 2006, year five in January 2010, etc.

Earnings. Earnings are total annual earnings for attached workers from all jobs, converted to 2016 dollars using the CPI-U.[4] There are two conditions that must be met by a graduate to be included in the earnings estimates for a given year. First, the graduate must have earned more than full time equivalent (35 hours a week for 50 weeks) at the prevailing federal minimum wage. Second, the graduate must have three or more quarters of non-zero earnings. These restrictions are in place so that our statistics capture the earnings of individuals who are reasonably attached to the labor market.

Employment. While earnings tabulations include earnings from all jobs, employment tabulations are based on the graduate's main job for that year. Main jobs are defined as the job for which graduates had the highest earnings in the reference year. To attach employer characteristics to that job, we assign industry and geography from the highest earnings quarter with that employer in the year. For multi-establishment firms, we use LEHD unit-to-worker imputations to assign establishments to jobs, and then in turn assign industry and geography. Graduates who fail to meet the labor force attachement restrictions described in the earnings section are categorized as non-employed, and assigned a firm industry and geography of "unclassified."

Tabulation Levels

Graduate Earnings

The Graduate Earnings tables, which summarize earnings outcomes for graduates, are at the Institution (6-digit OPEID), Degree Level, Degree Field, Graduation Cohort, and Year Post-Graduation level.

- **Degree Level**: We release data for the following degree levels:[5]
  - Associates;
  - Certificate, Less than 1 year;

[4] These earnings will be updated into current year dollars in future releases; however, all earnings will be stated in 2016 dollars before the protection system and then will be expressed in current year dollars.

[5] These degree levels were assigned according to the National Center for Education Statistics classifications.

  - Certificate, 1-2 years;
  - Certificate, 2-4 years;
  - Certificate, Post-Bacc
  - Certificate, Post-Masters;
  - Bachelors;
  - Masters;
  - Doctoral - Professional Practice;
  - Doctoral - Research/Scholarship.

- **Degree Field**: For the Certificate, Associates, Bachelors, and Doctoral - Professional Practice degree levels, the Degree Field is defined at the 4-digit CIP code level; for all other degree levels, the Degree Field is defined at the 2-digit CIP code level.
- **Graduation Cohorts**: For the Bachelors degree level, the graduation cohorts are three-year cohorts, e.g: 2001-2003; 2004-2006; 2007-2009; 2010-2012; 2013-2015. For all other degree levels, the graduation cohorts are five-year cohorts, e.g: 2001-2005; 2006-2010; 2011-2015.
- **Year Post-Graduation**: Earnings Outcomes and Employment Flows are reported 1, 5, and 10 years after graduation.

Employment Flows

The Employment Flows tables are counts of graduates (tabulated at the Institution, Degree Level, Degree Field, Graduation Cohort, and Year Post-Graduation level) and destination jobs (tabulated at the Firm Industry Sector and Geography level). The list below outlines values that are different from the Graduate Earnings tables above.

- **Degree Field**: The Degree Field is defined at the 2-digit CIP level for all degree levels.
- **Firm Industry**: Firm Industry is reported at the industry sector level. We assign firm industry as unclassified ("ZZ") for graduates classified as non-employed.

- **Firm Geography**: Firm Geography is reported at the Census Division level. Additionally, we release data on in-state flows for the state in which the institution is located.[6] We assign firm geography as unclassified ("Z") for graduates who are non-employed.

For workers that are not employed in a year, or do not meet the minimum earnings threshold we describe above, we assign them an industry sector 'Unclassified' and a firm geography of 'Unclassified.' We include this residual so that the rates from the Employment Flows have the same denominator across years post-graduation.

Dissemination

PSEO data are made available in raw form on the LEHD website `lehd.ces.census.gov/data/#pseo`, both in CSV and XLS formats. We have also developed a data visualization tool, to make the data more accessible for data users, which is available here: `https://lehd.ces.census.gov/data/pseo_beta_viz.html`.

Public-Use Data Files

We release two files, Earnings Outcomes (PSEOE) and Employment Flows (PSEOF). We release these files in CSV on our PSEO website. We provide theese files for all states together (all), as well as disaggregated by state of institution. For the state-level files, we also provide an additional XLS format file which has variable labels, and the PSEOE and PSEOF data in separate sheets. Because of the size of the PSEOF file, we do not report all aggregation levels in the XLS file, because in some cases it exceeds the number of allowable rows.

Updates and Future Planned Releases

PSEO data will be updated as new cells become available to publish. Since there are some years in which new cells are not available, this will happen almost annually.

[6] Reporting employment flows at the state level is under consideration for a future data release.

Comparability to Other Data

The College Scorecard is a data product released by the U.S. Department of Education beginning in 2013 and focuses on entering cohorts of students and their earnings ten years after initial enrollment, although they report longer-term outcomes as well. The Department of Education produces this product by matching federal financial aid data to IRS tax records. They report the 10th, 25th, 50th, 75th, and 90th percentiles of earnings for students. College Scorecard has recently included program-level data for very recent cohorts of graduates. College Scorecard differs from PSEO in source data, coverage of students and institutions, degree levels, employment measures, and privacy protection.

PSEO covers all graduates with matching UI earnings data from a smaller number of participating institutions, while College Scorecard provides information on a more comprehensive set of institutions, but only covers students who received Title IV federal aid. Earnings are reported in both products only for employed graduates (completers of a certificate or degree), but definitions of employment vary between the two products. PSEO reports earnings for graduates with at least full-time earnings at the federal minimum wage and at least 3 quarters of positive earnings in a measurement year. College Scorecard reports earnings for graduates with any earnings in a calendar year, but excludes those with loan deferments due to enrollment in schooling, military service, or if deceased prior to the end of the year. College Scorecard includes earnings for the self-employed, while PSEO does not account for self-employment earnings. The earliest cohort of graduates at the program level in College Scorecard is the 2014-2015 academic year. Earnings are thus currently reported 1 year after graduation. PSEO contains earlier cohorts of graduates back to 2001, and provides earnings 1, 5, and 10 years after graduation where available.

College Scorecard reports earnings at similar degree level categories and similar fields of degree as PSEO, with some slight differences. PSEO and College Scorecard degree levels coincide except for certificates. College Scorecard combines all undergraduate certificates of at least 1 year, while PSEO reports undergraduate certificates in less than one year, 1-2 year, and 2-4 year categories. Cohorts in College Scorecard are aggregated to 2 consecutive academic years of graduates, while PSEO cohorts are 3 calendar year graduate cohorts for baccalaureate and 5 calendar year cohorts for all other degrees. College Scorecard reports field of degree for all degree levels at the 4 digit

CIP code level, similar to PSEO for most degrees. PSEO reports field of degree at the 2 digit CIP code level for Masters and Doctoral-Research degrees. College Scorecard provides additional data on student debt and borrowing. PSEO provides information on geographic Census region, in-state counts, and 2 digit NAICS sector of employment in its Flows product. To protect the privacy of individuals in the data, College Scorecard suppresses cells with insufficient observations. PSEO uses a formally private mechanism and suppresses small cells due to data accuracy concerns.

Additionally, a number of states have released similar tabulations of graduate earnings by matching graduate records to in-state unemployment insurance records. While this match allows them to measure the earnings of graduates that stay within the state, these estimates are biased downwards, as mobility and higher wages are positively correlated.[7]

Protection System

One of the fundamental differences between the PSEO and previous data products released by LEHD is that outside parties have access to much of the underlying microdata used in this analysis and can therefore infer earnings of individuals that leave the state. Additionally, many of these states and systems have released earnings data from these matches and these data releases (and future possible data releases) must be considered public knowledge from the perspective of the Census. Finally, current LEHD protection systems are at the job level, but the frame of the earnings estimates is at the person level. For these reasons, we have to use differential privacy techniques to protect the data release. We describe these methods briefly below; for more detail, consult the appendix or Foote, Machanavajjhala and McKinney (2019).

Graduate Earnings Tabulations

In the Graduate Earnings tabulations, we release three percentile values (25th, 50th and 75th) and a cell count. To protect the earnings percentiles for a given cell, we categorize the earnings of all individuals into pre-defined histogram bins. The count in bin j of the histogram is characterized as q_j^c.

[7] A number of states have released estimates using in-state UI earnings: Colorado, Texas, and North Carolina, to name a few.

We then add noise to each bin according to a geometric mechanism, such that the protected count of individuals in a bin is $\tilde{q}_j^c = q_j^c + \zeta$, where ζ is the noise.[8] Adding noise to each bin count in the histogram means the entire list of protected bin counts is differentially private, which means any function of the counts is also differentially private.[9]

We use these counts to construct an empirical CDF, from which we calculate the percentiles. We also calculate the protected cell count from the sum of the bin counts. Cells with protected counts of less than 30 are suppressed and flagged.

Employment Flows Tabulations

For each count of graduates going from institution c, degree level l, degree field f, and graduate cohort g, to employment state s and industry k, T years after graduation ($Flow_{(clfg,ks),T}$), we use the geometric mechanism to add noise, as above.[10] Because we draw noise independently for each year post-graduation, the totals will not match across year post-graduation. We protect the employment flows to a state-sector and then aggregate these counts to the Census Division-sector level for publication.[11]

[8] Because all of these counts are integers, we use the geometric mechanism, which is the integer analog to Laplace noise. Formally, $\eta \sim X - Y$, where $X, Y \sim Geo(p), p = 1 - \frac{1}{e^\epsilon}$.

[9] We draw noise with the privacy loss parameter $\epsilon = 1.5$.

[10] We draw noise with the privacy loss parameter $\epsilon = 1.5$.

[11] We protect the flows at the state level so that if we want to release the flows at the state level at a future date, we can without incurring any additional privacy loss.

References

[1] John M. Abowd, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In *Producer Dynamics: New Evidence from Micro Data*, NBER Chapters, pages 149–230. National Bureau of Economic Research, Inc, september 2009.

[2] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially-private queries through consistency. *CoRR*, abs/0904.0942, 2009.

A Appendix on Protection System

This appendix details how we protect the released data. First, we discuss the Laplace noise infusion, which is used to protect the count queries for Graduate Earnings and Pipeline Flows. Second, we detail how we use Laplace noise in the histogram approach to calculate percentiles of earnings within a cell.

A.1 Protecting Count Data: Laplacian Noise Infusion

Consider a dataset d, and a neighboring dataset d' which differs by one observation. Furthermore, consider a count query $q_c(.)$ on a dataset, which returns the number of observations with certain attributes, which we will refer to as X. Now consider the cases below:

$$|q_c(d) - q_c(d')| = \begin{cases} 1, \text{if the differing observation has the attributes } X \\ 0, otherwise. \end{cases} \tag{1}$$

The sensitivity $S(q_c)$ is then the smallest number such that for any neighboring datasets,

$$|q_c(d) - q_c(d')| \leq S(q_c)$$

In the case of the count query, $S(q_c) = 1$. Therefore, for any count query $q_c(d)$, $q_c(d) + \zeta$ is ϵ-differentially private, where $\zeta \sim X - Y$, where $X, Y \sim Geometric(p)$ where $p = 1 - \frac{1}{e^\epsilon}$. For these tables, all the counts are integers, and therefore we will draw noise using the geometric noise mechanism.

A.2 Protecting Percentiles: Histogram Approach

To protect percentiles, we use an approach called the histogram approach.

Consider a dataset d, which has sorted values $e_1, e_2, ..., e_N$ and a function $H(.)$ that assigns each value e_i into a bin, grouping them together. Formally, this function is defined as:

$$H(e_i) = \begin{cases} 1, & \text{if } e_i \in [b_1, b_2) \\ & \dots \\ j, & \text{if } e_i \in [b_j, b_{j+1}) \\ & \dots \\ M-1, & \text{if } e_i \in [b_{M-1}, b_M) \\ M, & \text{if } e_i \geq b_M \end{cases} \tag{2}$$

Where the borders of the histogram bins, b_i, are public information. The key decision in implementing this protection method is determining how to set the bin definitions, which we describe in the next subsection.

Choosing Bin Definitions

We use the following bin cutoffs for the PSEO protection system. The bottom cutoff is \$10,000, which is very close to the minimum value in the data by construction, given that we restrict the sample based on earnings. For the next 19 b_is, we choose every 5th percentile of the log normal distribution with mean 11.003 and standard deviation 0.753.[12] Additionally, for b_M, we use the 97.5th percentile value of the distribution, which is about \$260,000. Finally, for any earnings greater than that value, we count it in the final bin, M. Together, we have 21 bins.

For reference, these histogram values are in the appendix.

Queries to Protect

From the definition of the histogram function above, the set of queries we protect are of the form q_j^c, which returns the count of the observations in a given bin j. Additionally, these queries imply the corresponding empirical CDF:

$$F(j) = \frac{\sum_{i=1}^{j} q_i^c}{\sum_{i=1}^{M} q_i^c} \tag{3}$$

[12]The log normal distribution is a good approximation of the overall earnings distribution. The mean and standard deviation were calculated using the 5-year ACS Public-Use Microsample. We calculated the mean and standard deviation of wage and salary income for employed individuals with a BA or above.

The sensitivity of each of these queries is 1, and therefore we can protect each of these queries with privacy loss ϵ by adding geometric noise as described above in Section A.1. Therefore, our protected counts are:

$$\tilde{q}_j^c = q_j^c + \zeta$$

Where ζ is drawn from a geometric noise distribution. The resulting histogram list of counts is ϵ-differentially private (Proposition 1 in [2]), and any function of these counts is also ϵ-differentially private because of the composition properties of differential privacy.[13]

Calculating Protected Percentiles

We use these fuzzed values to create a fuzzed CDF,

$$\tilde{F}(j) = \frac{\sum_{i=1}^{j} \tilde{q}_i^c}{\sum_{i=1}^{M} \tilde{q}_i^c}$$

If we assume that earnings are distributed uniformly within a bin, we can use $\tilde{F}(j)$ to extract protected percentiles.[14]

To calculate a percentile Y, suppose that it is in bin J such that

$$\frac{\sum_{i=1}^{J-1} \tilde{q}_i^c}{\sum_{i=1}^{M} \tilde{q}_i^c} < Y/100 \leq \frac{\sum_{i=1}^{J} \tilde{q}_i^c}{\sum_{i=1}^{M} \tilde{q}_i^c} \tag{4}$$

Then, the Yth percentile is $b_J + (b_{J+1} - b_J) \times \frac{(Y/100 \times \sum^{J} \tilde{q}_i^c) - \sum^{J-1} \tilde{q}_i^c}{\tilde{q}_J^c}$.[15] In the case where a percentile is in the largest bin, we define b_{M+1} to be the 99.9th percentile of earnings from the log normal distribution, which is 614597.[16]

We use this technique to calculate the 25th, 50th, and 75th percentile values.

[13] This result from Hay et al. (2009) allows the list of values $(\tilde{q}_1, \tilde{q}_2, ..., \tilde{q}_M)$ to also be considered releasable.

[14] Note that $\tilde{F}(j)$ will not necessarily be a true CDF, because there may be cases when $\tilde{q}_j < 0$.

[15] In words, if a bin J includes the Yth percentile, and the Yth percentile is W of the way through the interval defined by bin J, then the Yth percentile is the lower-bound value of bin J, b_J, plus $W \times width$.

[16] We calculate the percentile for the smallest bin such that Equation 4 is satisfied. This addresses the issue of negative counts in a subsequent bin, since it is possible for Equation 4 to be fulfilled in two distinct bins if the intervening bins have negative counts.

Calculating Protected Counts

We use the fuzzed counts from the histogram approach to calculate the total cell count. In this application, this is just a sum of all the histogram counts: $cellcount_c = \sum_{i=1}^{M} \tilde{q}_i^c$.

We determine whether or not to release values for a cell based on the total protected cell count, because values are noisier for smaller cells. Specifically, we do not release any data for cells with protected counts below 30. When we release tables, we will simply indicate that the cell count is below 30 and publish missing values.

A.3 Protecting Counts in a Sparse Matrix: Employment Flows

The EF data product is a set of count queries, and we will protect these queries using the geometric mechanism for noise.

Consider a flow ($Flow_{(clfg,ks),T}$) from institution c, degree level l, degree field f, and graduation cohort g, into industry k and state s, T years after graduation. To protect the flow count, we use the geometric noise mechanism to add noise to the count such that the protected count is:

$$\widetilde{Flow}_{(clfg,ks),T} = Flow_{(clfg,ks),T} + \eta$$

where $\eta \sim X - Y$, where $X, Y \sim Geometric(p)$ where $p = 1 - \frac{1}{e^{\epsilon}}$.[17] Therefore, $\widetilde{Flow}_{(clfg,ks),T}$ is ϵ-differentially private.

A.3.1 Post-Processing Protected Counts

Because the flows matrix is sparse, there will be a number of cases where $\widetilde{Flow}_{(clfg,ks),T} < 0$. While this flow is still ϵ-differentially private, it does not make sense logically from the perspective of a user since flows are strictly non-negative.

To that end, we post-process the differentially private counts in order to preserve the logic of the flows data. We post-process the flows in three steps, which we describe in detial. First, any flow that is negative is set to zero. Second, we calculate the difference between the new total count within a (c, g, f, d, T) cell before and after step 1. Finally, we adjust the counts of other non-zero

[17] In our setting, $\epsilon = 1.5$

cells so that the new total count within a (c, g, f, d, T) cell is the same. We describe this formally in the next sub-section.

A.3.2 Correcting Cell Counts

Because of the correction of negative flows above, as long as any flows are negative, $\sum_{k,s} \hat{Flow}_{(clfg,ks),T} > \sum_{k,s} \tilde{Flow}_{(clfg,ks),T}$, and therefore we need to correct the other counts in the matrix such that the overall total is unchanged. We do this correction by randomly selecting cells with non-zero counts in them, and subtracting, thereby adjusting to total count.

However, we do not want each cell's probability of being selected to be equal, because we know (from external sources) that some flows are less likely than others. We use three sources of data to weight a cell's probability of being selected for correction.

- State-to-State J2J Flows: We use the total accessions to destination states from an institution's state. We call this value $J_{h,d}$, for the flows from home state h to destination state d. This addresses the fact that some states are more connected than others; it is less likely for a graduate in Colorado to move to Maine than Arizona. We use these data from 2011-2015.

- Quarterly Census of Employment and Wages state-by-industry employment data: We use QCEW state-by-industry employment ($Emp_{s,k}$) to address the fact that within a state, some industries are smaller and therefore less likely to employ a graduate. For example, a flow into Information is more likely in California than a flow into mining. We use these data from 2011-2015.

- Institution by Field to Industry Sector Flows: Some institutions and majors have large flows to specific industry sectors, which may have small overall employment (e.g. Petroleum Engineering to Mining sector). We call these flows $INDFLOW_{c,f,k} = \sum_g \sum_d \sum_T \sum_s \widetilde{Flow}_{(c,g,f,d,T),k,s}$, and use the protected counts from above.

To weight the cells for selection in the below procedure, we multiply all of these values together and use the inverse, such that the $weight = \frac{1}{J_{j,d} Emp_{s,k} INDFLOW_{c,f,k}}$. In the case of any of these components being zero (or negative) we assign a value of 1. Using these weights, the algorithm for the correction is below.

1. Let $F = \sum_{k,s} \widehat{Flow}_{(clfg,ks),T} - \widetilde{Flow}_{(clfg,ks),T}$ which is number of jobs that need to be subtracted from the overall cell for the counts to be equal.

2. Randomly draw across the cells (using the weights above) with non-zero counts and each time a cell is drawn, subtract one from the flow count.

3. Repeat step 2 F times.

4. Recalculate the new flows $\overline{Flow}_{(clfg,ks),T}$ after these corrections.

By construction, $\sum_{k,s} \overline{Flow}_{(clfg,ks),T} = \sum_{k,s} \widetilde{Flow}_{(clfg,ks),T}$. These are the counts that are released to the public, as they satisfy the logical constraints on the data, while also being a consistent measure of the number of employed individuals.

A.3.3 Impossible Flows

There are some cells that are structurally zero, because we do not have data for those states and years. Below is a list of those states, and for which graduation cohorts and years after graduation they are removed.

- District of Columbia: 2001-2003, 1st year after graduation
- Massachusetts: 2001-2003, 1st and 5th years after graduation; 2001-2005, 1st year after graduation; 2004-2006 cohort, 1st year after graduation.

A.3.4 Suppression of Division Flows Data

Using the protected data at the state level, we will determine if missing data from a particular state causes an appreciable impact on the division-level flows that we report publicly. If we determine that an employment flow from a cell to a particular state is sufficiently large as a share of the flow to the division, we will suppress that cell if we are missing the state-level data. These suppression rules are similar to those implement in Job-to-Job Flows.

B Tables Appendix

Table 1: Histogram bin values

Bin	Lower Bound	Upper Bound
1	10000	17403
2	17403	22876
3	22876	27512
4	27512	31857
5	31857	36128
6	36128	40449
7	40449	44914
8	44914	49605
9	49605	54609
10	54609	60027
11	60027	65982
12	65982	72639
13	72639	80226
14	80226	89080
15	89080	99735
16	99735	113106
17	113106	130970
18	130970	157509
19	157509	207050
20	207050	262475
21	262475	614597

Notes: Except for the lowest value, these are all percentiles from a log normal distribution with mean 11.003 and standard deviation 0.753. Any observation will be classified into the final bin (21) if it has a value above 262475. For purposes of calculating the percentiles, we use the upper bound value for bin 21 of 614597, which is the 99.9th percentile of the log normal distribution.

Table 2: Coverages Dates by Post-Secondary System

System	Start Year	End Year	Number of Institutions
University of Texas System	2001	2016	15
Colorado Department of Higher Education	2001	2016	30
University of Michigan - Ann Arbor	2001	2016	1
University of Wisconsin - Madison	2001	2016	1