

NOISE INFUSION AS A CONFIDENTIALITY PROTECTION MEASURE FOR GRAPH-BASED STATISTICS

by

John M. Abowd*
Cornell University

Kevin L. McKinney*
U.S. Census Bureau

CES 14-30

September, 2014

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Fariha Kamal, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K132B, 4600 Silver Hill Road, Washington, DC 20233, CES.Papers.List@census.gov.

Abstract

We use the bipartite graph representation of longitudinally linked employer-employee data, and the associated projections onto the employer and employee nodes, respectively, to characterize the set of potential statistical summaries that the trusted custodian might produce. We consider noise infusion as the primary confidentiality protection method. We show that a relatively straightforward extension of the dynamic noise-infusion method used in the U.S. Census Bureau's Quarterly Workforce Indicators can be adapted to provide the same confidentiality guarantees for the graph-based statistics: all inputs have been modified by a minimum percentage deviation (i.e., no actual respondent data are used) and, as the number of entities contributing to a particular statistic increases, the accuracy of that statistic approaches the unprotected value. Our method also ensures that the protected statistics will be identical in all releases based on the same inputs.

*We acknowledge financial support from the U.S. Census Bureau and the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854-01, and the Alfred P. Sloan Foundation for LEHD infrastructure support. Abowd acknowledges additional funding through NSF Grants SES-0922005, SES-1042181, TC-1012593 and SES- 1131848.

1 Introduction

The advent of large-scale longitudinally linked employer-employee data, largely from administrative records, has the potential to greatly increase our understanding of the labor market. A prime example is the data developed by the U.S. Census Bureau's Longitudinal Employer Household Dynamics (LEHD) program. These new data allow researchers to follow employees over many years, capturing earnings at their initial as well as all subsequent covered employers. The observed employee mobility combined with a virtually universal for both employers and employees enables, for the first time, the estimation of labor market networks or graphs for an entire region or country.

Although these data are a rich new resource, privacy and confidentiality laws in countries such as the United States preclude statistical agencies from directly releasing detailed micro-level graph-based statistics. One solution to preserving respondent confidentiality while also preserving the analytical validity of released statistics is noise infusion. The technique of noise infusion is the primary disclosure avoidance mechanism used in the LEHD program's Quarterly Workforce Indicators (QWI).² Although not released as part of the QWI, a bipartite employer-employee graph underlies all of the published QWI statistics. In addition to the statistics already published, it is possible to project the employer-employee graph onto either the employer or the employee nodes, creating two related unipartite graphs. This paper describes our approach to expanding the existing LEHD noise infusion system to protect statistics based on these two projection graphs.

Graph Theoretic Representation of Linked Employer- 2 Employee Data

2.1 The Employer-Employee Bipartite Graph

Graph theory (see, *e.g.*, Diestel 2006) provides a way to organize and mathematically represent relationships between employees and employers. A graph $G = (V, E)$ consists of two sets—nodes V and edges E . Edges are created from the 2-element subsets of V , $E \subseteq [V]^2$. An edge (v_i, v_j) represents two nodes that are adjacent, *i.e.*, they have a direct connection (for example an employee works at an employer, friends in a social network, *etc.*). In real-world employer-employee data, for any given set of nodes V the number of realized edges is typically only a very small fraction of the total possible number of edges. For example, in the employer-employee graph each employee is typically only ever employed by a very small number of employers.

The nodes in the employer-employee graph can be separated into two distinct classes; employees $\{v_1, \dots, v_n\}$ and employers $\{v_{n+1}, \dots, v_m\}$.³ Thus there are n employees, $(m - n)$ employers, and m is the total number of both employees and employers.⁴ An edge is created when employee i is employed at employer j , defining a “job.” This relation is represented as a bipartite graph, where an edge $\{(v_i, v_j) \mid 1 \leq i \leq n; n + 1 \leq j \leq m\}$ can only occur between, not within, the two node subsets (employees and employers).

The set of edges can be represented in matrix form by the adjacency matrix A .

$$A = \begin{bmatrix} 0 & B \\ B' & 0 \end{bmatrix}$$

A is an $(m \times m)$ block diagonal matrix where $B = (b_{ij})_{n \times m-n}$ is known as the bi-adjacency matrix.

$$b_{ij} := \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

² See Abowd et al. (2009) and Abowd et al. (2012).

³ In the LEHD data there is no self-employment, employees do not employ other employees, and employers do not employ other employers.

⁴ In this paper we only consider the interesting case where both $n > 0$ and $(m - n) > 0$.

A specific example of B is given below for a labor market with 8 nodes (5 employees and 3 employers).

$$B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

The resulting graph for the 8-node adjacency matrix is shown in Figure 1 below.

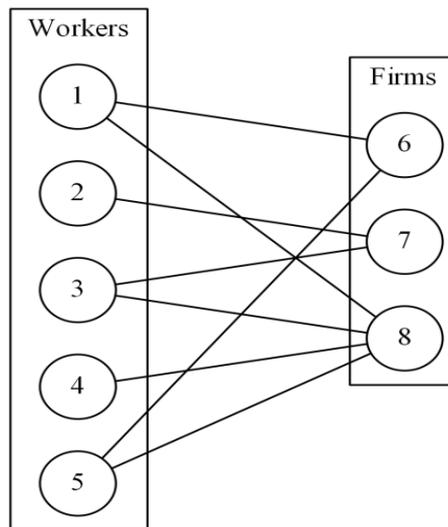


Figure 1 Example of the Employer-Employee Bipartite Graph

The straight lines between nodes represent the jobs (edges) and every edge in the graph corresponds to a non-zero element in the bi-adjacency matrix. Notice as well that the graph is bipartite, there are no edges within the class of employees or employers, only across classes. Another important property of the example graph is that all nodes are connected; there exists a path from any node to any other node.⁵

⁵ Most real-world data (LEHD data included) show a high level of connectedness; virtually all of the nodes (>95%) are in the largest connected component.



Figure 2 – An Employer-Employee Edge in Detail

Starting at the top of Figure 1, the first edge shows that employee one is employed by employer six for at least one quarter during the covered time period. In Figure 2 we present this job in more detail; showing the edge, nodes, and labels. From a graph-theoretic point of view, the presence of an edge represents the existence of some type of relationship between two nodes, while the characteristics of this relationship, e.g., the sequence of reported quarterly earnings, is called an edge label. The bipartite employer-employee graph also has labels for its nodes, which are independent of the existence of a particular employee-employer relationship. For example, an employee's sex or an employer's industry do not depend on whether an employee is employed by a particular employer at a particular point in time. The nodes, edges, and the labels represent the set of information we have about the employer-employee graph.

2.2 The Employer to Employer and Employee to Employee Graphs

The bi-adjacency matrix B of the employer-employee graph has special importance. It can be used to calculate information about the labor market such as the employment at each employer, S^F , and the number of employers for each employee, S^W :

$$S^F = B' * J$$

$$S^W = B * J,$$

where J is a conformable vector of ones.

The bi-adjacency matrix can also be used to better understand employee to employee or employer to employer networks. The projection of the employer-employee graph onto the employer nodes shows how employers are connected by employee mobility, while the projection onto the employee nodes shows how employees are connected through common employers. For example, edges in the employee to employee graph are formed when at least two employees are employed at the same employer, while both employee to employee and employer to employer edges are formed when an employee has multiple employers (not necessarily during the same time period). These types of relationships can more easily be understood by projecting the employer-employee

graph onto either the employee or the employer nodes. The correct projection matrix is simply the transpose of the bi-adjacency matrix itself.⁶

$$P^F = B' * B$$

$$P^W = B * B'$$

For the B in the example above, P^F and P^W are:

$$P^F = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & 1 \\ 2 & 1 & 4 \end{bmatrix}$$

$$P^W = \begin{bmatrix} 2 & 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 2 \end{bmatrix}$$

The projection onto the employer nodes, P^F , shows the number of employees at each employer on the main diagonal, while the off-diagonal elements show the number of employees employed at both employer i and employer j . The projection onto the employee nodes, P^W , shows the number of employers for each employee on the main diagonal, while the off-diagonal elements show the number of common employers for employee i and employee j .

The adjacency matrices for the employee and the employer graphs can easily be recovered from P by using the following formulas, where $I()$ is the indicator function applied to each element.

$$A^F = I(P^F - \text{diag}(S^F))$$

$$A^W = I(P^W - \text{diag}(S^W))$$

For the B in the example above, the resulting adjacency matrix A^F for the employer graph is:

$$A^F = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Notice that although B is not symmetric, the projections onto the employee and employer nodes, as well as the resulting adjacency matrices, are symmetric. Depending on the problem, it may be preferable to work with either P or A . The matrix P contains complete information about a particular graph, but the adjacency matrix A ignores any loops and/or multiplicity in P . In the employer to employer graph, employees who have only one employer create a loop (employee two and employee four in the example). If the loops are removed, only employees employed at more than one employer contribute to the set of edges. For the employee to employee graph loops are created by employers

⁶ This is not the usual projection matrix from linear algebra; it is a unipartite projection matrix.

that have only one employee (none exist in the example). In both graphs, multiplicity arises when the same two nodes are connected by multiple edges. Multiplicity is present in the example graphs whenever an off-diagonal element of P is greater than one or when a node has multiple loops.

A picture of the adjacency matrix for the employer graph is shown below in Figure 3.

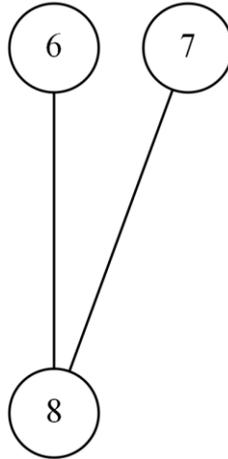


Figure 3 Adjacency Matrix for the Example Employer Graph

For the B in the example, the resulting adjacency matrix, A^W , for the employee graph is:

$$A^W = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Once again, the matrix is symmetric, and an employee must have been employed at an employer with at least one other employee to appear in the graph. In the example, all employees work at employers with at least one co-worker, but if a person were the only employee at an employer, and never worked at another employer with at least one other employee, that person would not be connected to any other employees (a row and column of the adjacency matrix would contain all zeroes). At least some mobility or multiple job holding is required for both the employer to employer and the employee to employer graphs to be connected. Without multiple job holding the set of edges in the employer to employer graph would be empty, while the employee to employer graph would contain isolated islands of edges, where the islands are made up of employees employed at the same employer. In the example, the employee-employer graph is connected, resulting in connected employer to employer and employee to employer graphs. A picture of the resulting example employee to employer graph is shown in Figure 4.

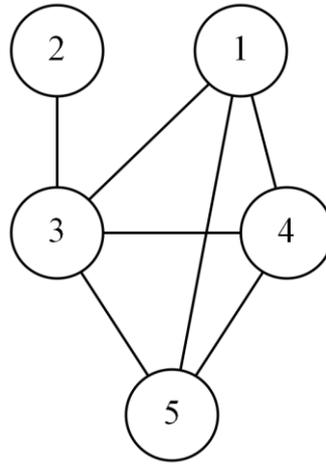


Figure 4 Example of the Employee to Employee Graph

The degree is the sum of the number of nodes attached to the edges of a given node. For example, in the employer to employer graph (excluding loops and multiplicity), the degree of nodes six and seven is one, while the degree of node eight is two. The two edges were created by both employee one and employee five being employed at employers six and eight (multiplicity of two), while employee three was employed at employer seven and eight. An analogous approach applies for the employee to employee graph except that the edges are created by employees employed at the same employer. Employers six and seven only have two employees and thus create only one edge each for the employee to employee graph. However, employer eight has four employees, resulting in six edges.

An employee with a large number of employers will generate a disproportionate number of edges in the employer to employer graph, while in the employee to employee graph a large employer will generate a disproportionate number of employee to employee edges. In both cases, the number of new edges, not counting the nodes that generate loops, is calculated by the following formula: $e = (z * (z - 1)) / 2$. For the employer to employer graph, z would be replaced with the number of jobs per employee, while for the employee to employee graph, z would be replaced with the number of employees at the employer.

3 The LEHD Infrastructure Data

At their core, the data in the LEHD infrastructure consist of three tables; a table of jobs (the EHF or Employment History File), a table of employee characteristics (the ICF or the Individual Characteristics File) and a table of employer characteristics (the ECF or the Employer Characteristics File). The EHF is built up from state level employer Unemployment Insurance (UI) reports of the quarterly earnings for all covered employees. Each record contains a unique employee identifier, a unique within state employer identifier, and the employee's quarterly earnings. The edges in the employee-

employer graph are precisely the jobs in the EHF. The employer to employer graph and the employee to employee graph are not directly stored by LEHD, but must be constructed using the job information in the EHF.

The ICF contains the labels for the employee nodes and the ECF contains the labels for the employer nodes. The employee labels are created at LEHD using internal Census databases. The ICF contains non-time varying employee characteristics such as sex, race, birth date, ethnicity, and completed education. The employer and establishment node data are collected by the participating states and forwarded to LEHD for incorporation into the ECF. The LEHD program combines these data with Census-derived location information to create the final ECF. The final file contains quarterly reports on industry, location, multiunit status, and size for both employer and establishment nodes.⁷

4 Calculating Graph-based Statistics

The graph-based statistics we discuss in this paper involve calculating sums of edges that meet a given set of selection criteria. The selection criteria are based on specific combinations of the labels associated with the edge itself and the edge's two nodes. To better understand how this works from a graph-theoretic viewpoint, we walk through the calculation of several statistics from the Quarterly Workforce Indicators (QWI). To show that the procedure is fundamentally similar when using one of the projections, we also present an example using a forthcoming LEHD product, the Job to Job Flows (JJF).

The employee employer graph contains the set of edges E generated in the labor market over a specific time period. Each edge in the set E represents a specific realized employment relationship between an employee and an employer. To make notation easier, we create another set $S = \{s | s \in \mathbb{N}, s \leq |E|\}$ along with a function $g: E \rightarrow S$. The function g maps every element of E to the sequential index S of the same size, allowing us to refer to each edge by number.

While calculating statistics on the entire graph is useful, it is arguably more interesting to compare groups over time or across some other characteristics. A variety of statistics can be produced by simply summing over an edge label L for a specific set of edges K , where $K \subseteq S$.

$$z = \sum_{s \in K} L(s)$$

⁷For more detailed information about the construction of the LEHD data see Abowd et al. (2009). Each state reports the employment relationship (job) at the level of the employer. For employers with multiple establishments, the multiple worksite imputation can be used to replace the employer node identifier with an establishment identifier. This allows characteristics of the establishment to be associated with a particular place of work. However, in either case the nature of the results we present are the same, thus we focus only on the employer nodes.

For example, to count the number of edges in the employee employer graph, define $L(s) = 1$ and $K = S$. A more realistic example would be to calculate beginning of period employment in 2005:1 for employees aged 18-24 employed by retail employers in Idaho. This calculation can be done two ways. In the first approach, set $L(s) = 1$ and define the set of edges to sum over as

$$K = \left\{ s \left| \begin{array}{l} s \in S, 18 \leq \text{age} \leq 24, \\ \text{industry} = \text{retail}, \text{firm location} = \text{Idaho}, \\ \text{earn}_{1994:4} > 0 \text{ and } \text{earn}_{1995:1} > 0 \end{array} \right. \right\}$$

In the second approach, define the edge label as

$$L(s) = \begin{cases} 1, & \text{if } \text{earn}_{1994:4} > 0 \text{ and } \text{earn}_{1995:1} > 0 \\ 0, & \text{otherwise} \end{cases}$$

and the set of edges to sum over as

$$K = \left\{ s \left| \begin{array}{l} s \in S, 18 \leq \text{age} \leq 24, \\ \text{industry} = \text{retail}, \text{firm location} = \text{Idaho} \end{array} \right. \right\}$$

In the first two examples, the edge label is binary, however, in some cases an edge label, such as earnings, must be used directly in the calculation. To calculate earnings for the same set of edges K , define the edge label as

$$L(s) = \begin{cases} \text{earn}_{1995}, & \text{if } \text{earn}_{1994:4} > 0 \text{ and } \text{earn}_{1995:1} > 0 \\ 0, & \text{otherwise} \end{cases}$$

The calculations are fundamentally similar when using a projection of the employee-employer graph onto the employer nodes. Instead of an employee and employer node, the nodes at both ends of the edge are employers, while the employee and the characteristics of the jobs are edge labels. Although job flow statistics are actually based on a directed version of the employer to employer graph, the undirected graph can still be used. The direction of each edge is determined as part of the calculation of the statistic. For example, how many jobs flowed from retail to manufacturing in Idaho during 1995:1? First, determine whether an employee moved in either direction during 1995:1 (left one job and started another job during the same quarter) and second, determine the direction. Both labels would then be used to calculate the flow statistic. Of course, there will also be a complementary flow statistic going in the other direction from manufacturing to retail. The sum of both directions should be equal to the total activity on that edge during 1995:1.⁸

⁸ Not all employer to employer edges will be classified as a flow. For example, multiple job holding at the same time or a period spent outside the labor market between two jobs would both result in edges not classified as flows

5 Applying Noise Infusion

Noise infusion guarantees that a predetermined minimum level of noise or distortion is applied to each data point, while also allowing the noise-infused statistic to approach the confidential value as the number of unique fuzz factors used to calculate the statistic increases. The fuzz factors for a specific employer/establishment are drawn from the ramp distribution $p(\delta)$, shown below and illustrated in Figure 5.

$$p(\delta) = \begin{cases} \frac{(b - \delta)}{(b - a)^2}, & \delta \in [a, b] \\ \frac{(b + \delta - 2)}{(b - a)^2}, & \delta \in [2 - b, 2 - a] \\ 0, & \text{otherwise} \end{cases}$$

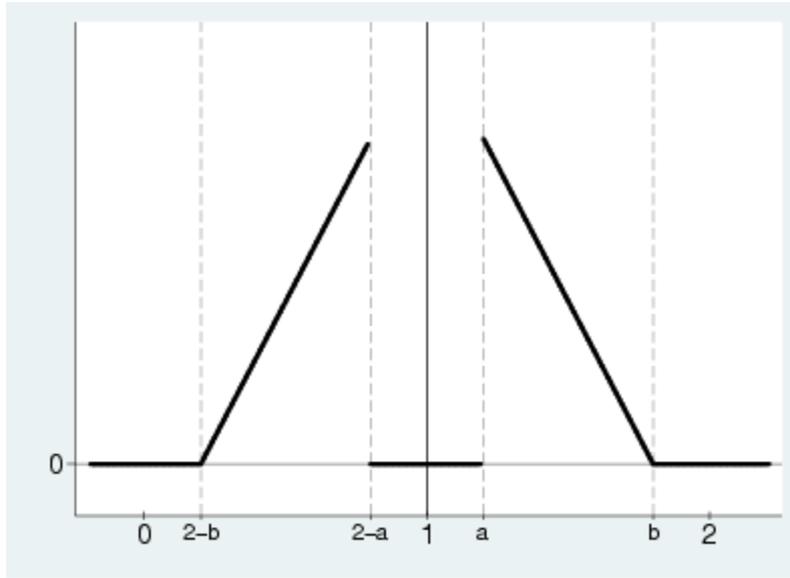


Figure 5 Graph of the Probability Density Function for the Fuzz Factor

The noise infused statistic z^* is calculated as follows.

$$z^* = \sum_{s \in K} \delta(s) L(s)$$

The fuzz factor can be viewed as another edge label, although the fuzz factor may not be unique for each edge. For example, all jobs at the same employer have the same δ .

For the employer-employee graph, convergence is a function of the number of unique fuzz factors (number of employers) used in the calculation. For a given number

of jobs, statistics composed of a large number of small employers should converge to the non-distorted value faster than one composed of a few large employers.⁹

Define the convergence ratio = [number of observations for a given population (employee, employer, job)]/(number of unique fuzz factors). When the number of unique fuzz factors equals the number of observations used to calculate a statistic the convergence is in some sense standard. When the ratio is greater than one the convergence is slower and when it is less than one it is faster. Given the design of the fuzz factors, the ratio will be greater than or equal to one for the employee employer graph because at least some employers will almost always have more than one employee. This is preferable from a disclosure avoidance perspective.

The employer to employer graph should also have a convergence ratio greater than one. Multiple employees often work at the same two employers, and for the edge to exist there will always be at least one employee and therefore one fuzz factor. Employers have edges with multiple employers, implying that the same fuzz factor could appear more than once. Once again this works toward increasing protection.

To protect statistics from the employee to employee graph, all edges created by employees working at the same employer must use the same fuzz factor. However, an employee can work with the same employer at two or more employers, creating multiplicity. In this case, which fuzz factor do we choose for that edge? One option is to choose an edge at random from among the N employers and use that fuzz factor. Some fuzz factors could get chosen multiple times, but this is not a problem from a disclosure perspective, all the fuzz factors provide protection. One complication in this case is that the fuzz factor for a given edge depends on the edge set, unlike in the employer to employer case where the fuzz factor chosen would always be the same for any edge between the same two employers.

6 An Employer Graph Example

6.1 Data

We demonstrate the performance of our noise infusion method using publicly available data on federal worker earnings histories provided by the United States Office of Personnel Management (OPM) in response to a Freedom of Information Act request from the Labor Dynamics Institute at Cornell University. The data contain information similar to those available in the LEHD infrastructure files, enabling the construction of an employer to employer graph. Although the data contain defense-related and overseas jobs, the employer information for these jobs is limited; therefore, we focus only on jobs at non-defense agencies located in the fifty states plus DC. Our analysis sample for years 2000-2012 contains 2,172,359 persons, 7,341 employers, 2,634,324 jobs, and 49,131,943 job-year-quarter earnings observations.

An employer in our sample is defined by the intersection of the state and the agency/sub-element. This definition results in state based employers, similar to the way

⁹⁹ This definition is exactly the method used in the Quarterly Workforce Indicators, a dynamic extension of Evans et al. (1998).

employers are defined in LEHD data.¹⁰ However, compared with data for the private sector, there are relatively few federal employers—only 7,341 compared with about 15 million in LEHD data over a similar time period. Given the relatively small number of employers per worker and the structure of the federal government employment relationship, it is perhaps not surprising that federal workers have relatively few federal jobs. About 85% of federal workers in our sample have only one employer, while in LEHD data, over a similar time period, about 20% of the workers have only one employer.

The relatively low level of worker mobility results in fewer edges in the employer projection of the employer-employee graph for a given size labor market and length of analysis period. For illustrating the performance of the noise infusion method this is a nice feature, but for applications with a higher proportion of small employers, the distortion of the typical statistic will be less than is shown in our example.

The employer graph constructed from the OPM data contains 123,930 employer-employer edges. Each edge also has an associated set of jobs/workers called the multiplicity. For loops, this number represents the sum of the workers at the firm with only one observed employer. In this case, the number of jobs and workers are equal. For edges that are not loops, each worker employed at both employer nodes at some time during, not necessarily contemporaneously, 2000 to 2012 contributes one to the value of the multiplicity for that edge. The total multiplicity in our employer graph is 2,465,461.

6.2 Analytical Validity Measures

Jensen-Shannon Distance

Let z and z^* be the unprotected and protected values of the statistic of interest. We assume that both statistics are job counts, as they are in the application below, but the formulas are valid for any magnitude measure. Assume, again as in the application below, that there is a mutually exclusive, exhaustive classification of all jobs in the universe for z indexed by $\ell = 1, \dots, L$. Define

$$\pi_\ell = \frac{z_\ell}{\sum_{s=1}^L z_s} \text{ and } \pi_\ell^* = \frac{z_\ell^*}{\sum_{s=1}^L z_s^*},$$

which expresses the magnitude measure as a fraction of the total for the universe—all jobs in the example below. Then, the Jensen-Shannon distance measure is defined as:

$$JSD(\pi, \pi^*) = \sqrt{\frac{1}{2} \sum_{\ell=1}^L \pi_\ell \log_2 \frac{\pi_\ell}{\left(\frac{1}{2}\pi_\ell + \frac{1}{2}\pi_\ell^*\right)} + \frac{1}{2} \sum_{\ell=1}^L \pi_\ell^* \log_2 \frac{\pi_\ell^*}{\left(\frac{1}{2}\pi_\ell + \frac{1}{2}\pi_\ell^*\right)}}.$$

¹⁰ Documentation for the OPM data along with a listing of the complete set of agency/sub-element codes is available in Office of Personnel Management 2014. A sub-element is an administrative division of the agency.

Root Integrated Mean Squared Error

Using the same inputs the root integrated mean squared error is defined as

$$RIMSE(\pi, \pi^*) = \sqrt{\sum_{\ell=1}^L (\pi_{\ell} - \pi_{\ell}^*)^2}.$$

6.3 Disclosure Protection of the State-State Employer Graph

Using the employer graph and applying our edge-based noise infusion method (see the technical appendix for more details) we produce two state-to-state mobility tables—one showing employer relationships and the second weighted by the number of employees associated with each employer-to-employer edge (multiplicity). We also produce the same two state-to-state mobility tables without applying our noise infusion method.¹¹ This allows us to compare each table (both with and without noise infusion) and assess the performance of our method.

The results are shown in Table 1 for the 1,326 cell state-to-state mobility table. To prepare the noise infused statistics, we used parameter values for the ramp distribution of $a = 1.15$ and $b = 1.25$.

Table 1 Summary Statistics for the OPM State-to-State Mobility Table

	Employer		Worker	
	Mean	StdDev	Mean	StdDev
Unique Edges in Cell				
1-6 (P5)	0.005	0.114	0.011	0.120
7-9 (P10)	0.009	0.060	0.026	0.083
10-19 (P25)	0.006	0.055	0.015	0.080
20-35 (P50)	0.003	0.043	0.007	0.068
36-75 (P75)	0.006	0.033	0.005	0.061
76-161 (P90)	0.004	0.026	0.004	0.055
162-272 (P95)	0.001	0.026	0.001	0.055
272-15035 (P100)	0.004	0.021	0.010	0.057
JSD	0.011		0.027	
RIMSE	0.002		0.010	

Table 1 shows the performance of the noise infusion method for various cell size classes. For example, the first row shows results for cells with one to six employer-employer edges, where six is also the fifth percentile of the cell size distribution. These are cells where relatively few workers were employed in both states, for example AK and DE. The last two rows show the overall analytical validity statistics, JSD and RIMSE. Both statistics overall statistics may be interpreted as average percentage point

¹¹ The two tables, both with and without noise infusion, are available upon request.

discrepancies between the fuzzed and underlying tables. In the Employer column the JSD of 0.011 indicates an average discrepancy of 11 basis points (27 basis points in the Worker weighted variant). The RIMSE estimates are of a similarly small magnitude.

For each cell in the table, we also calculate the statistic c defined as:

$$c = \frac{\pi_\ell - \pi_\ell^*}{\left(\frac{1}{2}\pi_\ell + \frac{1}{2}\pi_\ell^*\right)}.$$

The average and standard deviation of the statistic c for the employer and the worker (multiplicity) table are shown in columns two through five. In each of the size classes, the average percentage difference between the true and the noise infused value is close to zero (column two and four). On average, even in relatively small cells the noise infused values are centered on the truth, however the dispersion of the percentage difference is much higher for smaller cells. The dispersion is also higher in general for the worker (multiplicity) table, due to the large variance in multiplicity across employer-to-employer edges, thus disproportionately magnifying the effect of certain edge fuzz factors.

7 Conclusion

We have defined and implemented an extension of the dynamically-consistent noise infusion method for confidentiality protection originally used with the U.S. Census Bureau’s Quarterly Workforce Indicators. Our extension shows how to apply the method to graph-based statistics that are calculated from the employer projection of the basic employer-employee graph in the LEHD data. One, and only one, fuzz factor is used for each pair of employer’s contribution to the statistic. The correct fuzz factor is determined algorithmically to insure that the resulting noise-infused statistic inherits the analytical validity properties of the basic noise-infusion method. The confidentiality protection is insured by the design of the fuzz factors.

One thing we have not discussed in this paper is how to use this framework when calculating more complicated graph (edge) based statistics such as path length, centrality measures, etc. We think we should be able to apply a fuzz factor to each edge used in the calculation, but this may or may not make sense for some statistics.

If the matrix P^F is not complete, then additional measures must be taken prior to the release. The simplest approach is to suppress the cells that do not meet some minimum size criterion D . However, when a large number of cells in P^F lie in the interval $[0, D)$, the information loss of using suppression may be so large that other alternatives may be preferred. One alternative is to synthesize the cells that would otherwise be suppressed. This approach preserves a large portion of the information in P^F , especially the presence of the existence of an edge between two employers A^F .

References

1. Abowd, J, B. Stephens, L. Vilhuber, F. Andersson, K. McKinney, M. Roemer, and S. Woodcock (2009) “The LEHD Infrastructure Files and the Creation of the Quarterly Workforce

- Indicators” in T. Dunne, J.B. Jensen and M.J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data*, Chicago: University of Chicago Press for the National Bureau of Economic Research, pp. 149-230.
2. Abowd, J, K. Gittings, K. McKinney, B. Stephens, L. Vilhuber, and S. Woodcock (2012) “Dynamically Consistent Noise Infusion and Partially Synthetic Data As Confidentiality Protection Measures for Related Time-series,” Federal Committee on Statistical Methodology, Office of Management and Budget, 2012 Research Conference Papers, available at http://www.fcsml.gov/12papers/Vilhuber_2012FCSM_VIII-C.pdf (cited May 18, 2014).
 3. Diestel, R., (2006) *Graph Theory*, Berlin, Springer-Verlag.
 4. Evans, T., Zayatz, L., and Slanta, J. (1998) “Using noise for disclosure limitation of establishment tabular data,” *Journal of Official Statistics* 14, 537-551.
 5. Office of Personnel Management (OPM). February 28, 2014, “The Guide to Data Standards Part A: Human Resources,” retrieved from <https://www.opm.gov/policy-data-oversight/data-analysis-documentation/data-policy-guidance/#url=Data-Standards> on April 15, 2014.

Technical Appendix

Our employer graph noise infusion algorithm exploits the existing fuzz factors available for each employer in the LEHD data. To create the fuzz factor for an employer-to-employer edge, we randomly select one fuzz factor from the two available, designating the chosen employer fuzz factor as the new edge fuzz factor. The new edge fuzz factor is then be used to multiplicatively modify every value in all subsequent statistics and tabulations.

To reduce the variability across releases of the same table, each fuzz factor should, whenever possible, be chosen according to a deterministic rule that exploits any randomness present in the digits of the employer identifier. In the LEHD data two digits are extracted from each twelve digit SEIN starting at position eight.¹² The following decision rule is used to select the fuzz factor. If the two digits extracted from the first SEIN are less than the two digits extracted from the second SEIN, then the edge is assigned the fuzz factor from the first SEIN. If the converse is true then the fuzz factor for the second SEIN is used. If a tie is encountered, the tie is resolved using the first digit of the two-digit extract (position eight in either SEIN). If the first-digit value is a zero or below (where negative values come from the ASCII codes for characters, which occur occasionally in the SEINs), then the edge receives the fuzz factor associated with the first SEIN; equal to one, then the edge receives the fuzz factor associated with the second SEIN; ... ; equal to eight, then the edge receives the fuzz factor associated with the first SEIN; and a nine or above, then the edge receives the fuzz factor associated with the second SEIN.

For the data used in the example, the digits of the employer ID (agency/sub-element) were not amenable to the deterministic rule, therefore one of the two employer node fuzz factors was chosen at random and that value was assigned to be the fuzz factor for the edge.

¹² The last two digits of the SEIN had a very uneven distribution (too many zeros, for example), but the two digits starting at position eight have an empirical uniform distribution. To get around the occasional non-numeric character, the SEIN is not convert to numeric and the inequality comparisons are made using the ASCII codes.