Statistical Journal of the IAOS 32 (2016) 81–92
DOI 10.3233/SJI-160962
IOS Press

Differentially private publication of data on wages and job mobility

Ian M. Schmutte
Department of Economics, Terry College of Business, University of Georgia, Athens, Georgia
E-mail: schmutte@uga.edu

AUTHOR COPY

Abstract. Brazil, like many countries, is reluctant to publish business-level data, because of legitimate concerns about the establishments' confidentiality. A trusted data curator can increase the utility of data, while managing the risk to establishments, either by releasing synthetic data, or by infusing noise into published statistics. This paper evaluates the application of a differentially private mechanism to publish statistics on wages and job mobility computed from Brazilian employer-employee matched data. The publication mechanism can result in both the publication of specific statistics as well as the generation of synthetic data. I find that the tradeoff between the privacy guaranteed to individuals in the data, and the accuracy of published statistics, is potentially much better that the worst-case theoretical accuracy guarantee. However, the synthetic data fare quite poorly in analyses that are outside the set of queries to which it was trained. Note that this article only explores and characterizes the feasibility of these publication strategies, and will not directly result in the publication of any data.

Keywords: Demand for public statistics, technology for statistical agencies, optimal data accuracy, optimal confidentiality protection, matched employer-employee data, job mobility, differential privacy

1. Introduction

Statistical agencies exist to collect and publish social and economic data. In doing so, they face a dual mandate to make published information as accurate as possible while protecting the privacy and confidentiality of individuals and businesses. Ideally, a statistical agency can formalize the technological tradeoff between privacy and accuracy: "How much privacy loss must be incurred to increase statistical accuracy by a given amount"? Such formalizations are not easy to provide, and seldom result in a crisp articulation of what an economist would call the marginal cost of accuracy.

Very recent developments in formally private data publication strategies that yield explicit theoretical tradeoffs between privacy and accuracy have emerged from the cryptography literature. Some of the most promising of these, which I consider exclusively here, are organized around the concept of "differential privacy" [1,2]. Under differential privacy, privacy loss is quantified, roughly, as the maximum possible change in the posterior belief upon seeing the published data, of an attacker that a particular item was in the database. The key differential privacy parameter, ε, measures privacy loss – higher values correspond to a larger amount of information about a particular item being leaked through publication of statistics using the differentially private mechanism.

For most differentially private mechanisms, and specifically for the mechanism I consider here, it is possible to establish a lower bound on the accuracy of published statistics. The theoretical guarantee formalizes the intuition that accuracy is increasing in privacy loss. As privacy decreases (ε increases), accuracy increases. This feature of differentially private mechanisms make them particularly appealing because they describe the set of "production possibilities" afforded to a statistical agency or other data custodian in possession of a fixed database when using a particular differentially private data publication technology. In combination with information on social preferences for privacy relative to accuracy, it may be possible to determine the socially optimal choice of privacy and accu-

1874-7655/16/$35.00 © 2016 – IOS Press and the authors. All rights reserved

racy to provide. This possibility is considered in great detail by [3].

The goal of this paper is to assess the tradeoff between the privacy guarantee and statistical accuracy when applying a differentially private data publication technology to a real-world use case. Because of their computational complexity and limited domain of application, much of the research on differentially private algorithms is highly theoretical. As a result, it is difficult to assess how effective differential privacy can be as a general approach to data publication. This is unfortunate, both quantifying the privacy guarantee is potentially quite useful in a policy setting, and also because some differentially private mechanisms would allow publication of relational information and network statistics for which it is quite difficult to ensure privacy using existing approaches.

My application is to the publication of statistics from Brazilian matched employer-employee data. Specifically, I focus on data that characterize job-to-job mobility patterns. The analysis data is a set of records that each correspond to a new job. For each job, there are three discrete characteristics: (1) the employer-specific wage premium of the worker's new employer (10 categories, based on deciles of the estimated wage premium), (2) the employer-specific wage premium of the workers previous employer (10 categories, again, assigned as deciles, plus one category to capture workers who come from non-employment), (3) a match-specific contribution to wages (10 categories, based on deciles of the distribution of within-job wage residuals). The result is a dataset with 1,100 possible combinations of the three categorical variables, and the analysis is performed on the resulting contingency table (or histogram) representation of the data.

I proceed by generating differentially private synthetic data by implementing the MWEM (Multiplicative Weights – Exponential Mechanism) algorithm due to [4]. Their method operates by storing an approximation, A, of the true database. The approximation is updated adaptively to provide improved answers to a target set of analyses, or queries. Upon completion, the algorithm releases A, which is guaranteed to satisfy ε differential privacy and also to have a worst-case error that is asymptotically lower than the database size, and potentially much lower.

The application I consider is far from trivial. Statistics characterizing the mobility of workers between jobs and its association with components of earnings heterogeneity are central in labor market analyses using linked employer-employee data. [5] use these statistics to test the exogenous mobility assumption required for interpretation of the worker and firm components of earnings in models derived from [6]. [7] use related statistics similarly. [8] argues for the existence of "job ladder" behavior, in which workers tend to move into jobs with higher-paying employers. In each application, the analyses were conducted on restricted-access matched employer-employee data, and it would not be possible to replicate those analyses, not to consider extensions, without access to the underlying microdata. The ability of researchers to vet and extend the results of this research would be greatly facilitated by the production of synthetic data. Finally, the U.S. Census Bureau has recently started to produce statistics on job-to-job flows as part of the Quarterly Workforce Indicators (QWI) [9], for which there is already a considerable demand from academic researchers as well as from local and regional planners.

The results of this analysis indicate that the MWEM mechanism can be an effective tool for producing synthetic data on job mobility. For the set of queries used to train the algorithm, the synthetic data performs very well, and in fact gives worst-case error far lower than the theoretical bound. The empirical accuracy is still monotonic and convex in privacy loss, suggesting that the empirical tradeoff between privacy and accuracy could serve as the basis for a production possibilities frontier in an analysis of the optimal data publication strategy. On the negative side, the data fares rather poorly in addressing a set of queries for which it was not trained. These results indicate that the MWEM algorithm can be used to generate synthetic data, but care must be taken to consider in advance the type of analyses that are to be performed, or to reserve part of the privacy budget to allow for extensions of the set of possible analyses down the road.

2. Differential privacy and the multiplicative weights exponential mechanism (MWEM) algorithm

This section introduces the key concepts required to understand the MWEM algorithm and its associated privacy and accuracy guarantees. My discussion draws heavily on the descriptions and notation in [4].

2.1. Databases, histograms, and linear queries

The data custodian is in possession of a database, B, whose entries, i, are drawn from a domain, $D =$

AUTHOR COPY

$D_1 \times \ldots \times D_K$ where D_k is a finite discrete set for each k. The domain D has cardinality $|D| = \prod_k |D_k|$ Let $R = \{0,1\}^{|D|}$. There is a function, $h : D \to R$ that maps each element of the domain of D onto a basis vector from R. So, if there are n entries in B, then $B \in D^n$. We can define the dataset analogously as $h(B) \in R^n$ where we abuse notation to indicate the composition of h applied to each of the n elements in B.

In practice, we work with the histogram representation of B, which is defined as $H = \sum i = 1^n h(B) \in \{0, \ldots, n\}^{|D|}$. H is clearly a vector whose jth entry is the frequency of entries, i, satisfying $d_i = h^{-1}(r_j)$. From here out, we will work exclusively with histogram representations of the underlying databases, and use the notation H_j to refer to the jth entry. We may also abuse notation and use $H(d)$ to refer to the count of observations in the histogram that represent occurrences of $d \in D$. The distance between two histograms with the same cardinality, $|D|$ is easily represented by $||H-K|| = \sum_{j=1}^{|D|} |H_j - K_j|$. This measures the number of records that would have to be changed to turn H into K.

Analysis of the discrete data embodied in the histogram occurs through *linear queries*. A linear query simply counts the number of records that satisfy some linear combination of attribute settings. Therefore, the linear query can be represented by a vector, q, of length $|D|$. The answer to the query is $a = q' * H \in R$, where R is the range of the query space. For most of our analysis, I will consider queries whose entries are either zero or one. These suffice to generate all margins of the contingency table. However, the class of linear queries is broader.

2.2. Differential privacy

Differential privacy formalizes the idea that data publication should limit the amount of information that can be learned about any individual in the database. Data publication is modeled as a random algorithm, and privacy is achieved by guaranteeing that the mechanism will behave similarly on two databases that are nearly the same.

A key feature of differential privacy is that the privacy guarantee is a feature of the mechanism, but does not depend on the data. Furthermore, the privacy cannot be compromised through post-processing of the output. This is in part because the privacy guarantee also does not depend on whatever external information might be possessed by an outside analyst or attacker.

Definition 1. *(Differential Privacy) Let M be a random mechanism that maps histograms, H, to distributions over an output space, R. We say M provides (ε, δ)-differential privacy if for every $S \cup R$ and for all histograms H and K where $||H - K|| \leqslant 1$*

$$\Pr[M(H) \in S] \leqslant exp(\varepsilon) \Pr[M(K) \in S] + \delta. \quad (1)$$

Note that in the case $\delta = 0$, which I consider extensively in the application, the convention is to say that M satisfies $(\varepsilon, 0)$-differential or just ε-differential privacy.

2.3. The multiplicative weights exponential mechanism (MWEM)

[4] developed the MWEM as a simple mechanism for private query release. The algorithm has two key features. It is operated "offline", which means all queries are posed in advance. Second, it can be used either to publish responses to the queries in that set, or to publish a synthetic version of the raw input data. For either application, the privacy and accuracy guarantees are the same.

The key to the algorithm's efficiency is in iterating between updating the synthetic database that gives approximate answers to the chosen queries (the Multiplicative Weights [MW] step), and choosing a "good" query to use for the next update (the Exponential Mechanism [EM] step). The guarantee of ε-differential privacy arises from noise added during the MW step and also noise added via the Laplace Mechanism (described below) together with composition arguments that are standard in the differential privacy literature.

Relative to other differentially private algorithms, MWEM is useful because it has a worst-case error that is provably smaller than n, the total number of records. While obtaining this error guarantee is important theoretically, it establishes just that using MWEM is superior to publishing the total number of records, n, but nothing else. In light of this, it is meaningful to find that in our real-world use case, the actual error is much lower. The MWEM algorithm is also scalable to databases with a large number of attributes (for which $|D|$ is on the order of 2^{1000}). The application in this paper has $|D| = 1,100$, so is relatively small by this standard. Future research will consider applications that extend both the number of attributes and the total number of records in the database.

AUTHOR COPY

2.3.1. The algorithm

Algorithm *Multiplicative Weights Exponential Mechanism*

Input: Data set, H, over a universe, D; a set Q of linear queries; total number of iterations $T \in N$; privacy parameter $\varepsilon > 0$. The number of records in H is n.

1. Initialize the synthetic histogram, K_0, as n times the uniform distribution.
2. **for** $t \leftarrow 1$ **to** T
3. Exponential Mechanism Step: Select a query, $q_t \in Q$ using the Exponential Mechanism parameterized with $\varepsilon/2T$ and score function
$$s_t(H, q) = |q' K_{t-1} - q' H| \quad (2)$$
4. *Laplace Mechanism*: Set measurement $m_t = q_t' H + Lap(2T/\varepsilon)$.
5. *Multiplicative Weights Step*: Let K_t be n times the distribution whose entries satisfy
$$K_t \propto K_{t-1} \times exp(q_t \times (m_t - q_t' K_{t-1})/2n) \quad (3)$$
6. **Output:** K as the simple average across all K_t for $t < T$.

The reader is referred to [4] for details of each step. In implementing the algorithm, [4] find performance is improved if, at each iteration, t, the MW step is repeated for all previously selected queries. This can be done without reducing the privacy budget, so the only cost is in computation time. The increase in overall accuracy is considerable. For a large query set, though, the increase in computation time is not negligible.

2.3.2. Privacy and accuracy guarantees

Here, I simply repeat the privacy and accuracy guarantees proven by [4]. The key feature for this analysis is the worst-case accuracy guarantee, which tells us the worst-case error in response to any query. The theoretical worst-case accuracy guarantee is a general result. In the present application to RAIS data, the actual worst-case error is considerably lower.

Theorem 1. *The MWEM satisfies ε-differential privacy.*

Theorem 2. *Given any dataset, H, with n records, together with a set of queries, Q, number of iterations T, and $\varepsilon > 0$, with probability at least $q - 2T/|Q|$, MWEM produces synthetic histogram K that satisfies*

$$max_{q \in Q} |q' H - q' K| \leqslant 2n \sqrt{\frac{log|D|}{T}} + \frac{10 T log|Q|}{\varepsilon}. \quad (4)$$

When the number of iterations is chosen to minimize error, the worst-case error is bounded above by n.

3. Data

I use the MWEM mechanism to generate synthetic dat a on job-to-job mobility patterns in data from Brazil's *Relação Anual de Informações Sociais*, or Annual Social Information Survey (RAIS). I use the linked data to estimate a decomposition of log wages into parts associated with time-varying observables along with permanent worker and establishment heterogeneity. Next, I assign each observation into deciles of the worker and establishment effect distribution. I also compute the average residual within each match, and assign these to deciles as well. The final dataset of interest measures the frequency of transition between jobs in different deciles of the establishment effect distribution, disaggregated across deciles of the average residual (orthogonal match effect) distribution.

3.1. RAIS data

RAIS is a census of formal sector jobs. Each year, the Brazilian Ministry of Labor and Employment (MTE) collects data on every formal sector job for the purpose of administering the *Abono Salarial* – a constitutionally mandated annual bonus equivalent to one month's earnings. The information in RAIS is provided at the establishment level by a company administrator. In smaller firms and plants, this is likely the owner or plant manager; in larger establishments there may be dedicated personnel who submit the information. Coverage is universal, as employers who fail to complete the survey face mandatory fines and also risk litigation from employees who have not received their *Abono Salarial*.

For every job, the employer reports information on the characteristics of the worker, including a unique identifier that allows us to track the worker from job-to-job. The employer also reports information on the characteristics of the job, including the average monthly wage, the actual wage in December, the number of contracted hours per week, and the occupation. The employer also reports basic characteristics of the plant, including a common identifier and information on plant's industry, location, and the number of employees.

In Brazil a worker is formally employed if he or she has a registered identification number with one of

AUTHOR COPY

two social security programs: the *Programa de Integração Social* (PIS), or Social Integration Program, or the *Programa de Formação do Patrimônio do Servidor Público* (PASEP), or Civil Servants Equity Formation Program, depending on if the worker is employed in the private sector or the public sector. PIS/PASEP numbers are consistent across workers and follow a worker for life. For firms, formal employment means that the employer contributes to a bank account administered by either *Caixa Econômica Federal*, if registered with PIS, or *Banco do Brasil*, for PASEP workers, covering all worker categories. Formal employers must also have employment contracts for all employees.

3.2. Earnings decomposition

I borrow estimates performed by Lavetti and Schmutte (2015) of the canonical two-way decomposition of earnings heterogeneity [6] using the conjugate gradient method to obtain the full least-squares solution and applying identification following the methods described in [10]. The empirical model is

$$w_{it} = x_{it}\beta + \theta_i + \psi_{G(i,t)} + \varepsilon_{it}. \tag{5}$$

Here, a unique "employer", g, is defined by a combination of plant and occupation. The indicator function $G(i,t) = g$ if worker i was employed in plant-occupation combination g in year t, and $\psi_{G(i,t)}$ measures plant-occupation-specific variation in compensation . The dependent variable, w_{it}, is the log wage, and θ_i captures characteristics of the individual that do not change over time and are correlated with wages. As is common in the econometric literature, the model is estimated by fixed effects to allow arbitrary correlation between the time-varying observables in x, the worker effects, θ, and the plant-occupation effects, ψ.

Table 1 repeats statistics reported in Lavetti and Schmutte (2015) describing the mean, standard deviation, and correlations among the effects obtained when estimating Eq. (5). The statistics are computed by merging log wage components back to the estimation file, so these statistics are weighted by total employment and employment duration.

3.3. Earnings heterogeneity and job mobility

My goal is to characterize the relationship between job mobility and the employer-specific contributions to earnings. From the full analysis file, I restrict the sample to include just observations for worker-years in which the worker begins a new job. For each such observation, I record whether the worker was employed or unemployed prior to starting their new job. If they were employed, I record the identity of the previous employer (plant-occupation pair). More specifically, I restrict attention to job spells for each worker's dominant employer (defined as the employer for whom they worked the most hours that year). Then, if a worker starts a new dominant job, I record whether they are moving from non-employment or if they are moving to a new dominant employer.

To this sample of job transitions, I match the plant-occupation effect for each dominant employer together with the average residual across all periods for which the job is observed in the data. For job-to-job transitions, I also match the plant-occupation effect of the originating job. Next, I compute the deciles of the distribution of plant-occupation effects across all plant-occupations observed in the data. I also compute deciles of the distribution of match-specific average residuals (which I will refer to as match effects hereafter). Finally, I match each observation to the associated decile in the plant-effect distribution for the destination job and, if relevant, originating job. I also match each observation to its decile in the match effect distribution. The remainder of the analysis in this paper is based on a five percent simple random sample from the job transition data. The data set contains three categorical variables: the decile of the plant-occupation effect for the origin job (with a separate classification if the worker was not employed in the previous year), the decile of the plant occupation effect for the destination job, and the decile of the match effect. There are $11 \times 10 \times 10 = 1,100$ possible combinations in these data.

Figures 1 and 2 display some of the important second-order marginals that characterize job-to-job mobility. Figure 1 shows the distribution of new jobs across deciles of the employer-effect distribution when workers move from non-employment. The data suggest workers are equally likely to move into jobs at the second through tenth decile. Transitions from non-employment are much less likely to end in jobs with the worst-paying employers.

Figure 2 shows the distribution of destination jobs across all deciles of the employer effect distribution for transitions in which the worker was previously in a job with another employer. The distributions are shown conditional on the decile of the origin job. For example, in the figure, the closest ribbon shows the distribution of destination employer types among all jobs

AUTHOR COPY

Table 1
Correlation among components of the log wage rate: RAIS 2003–2010

	Mean	Std. Dev.	Correlation				
			Log wage	$X\beta$	θ	ψ	ε
Log Wage	1.30	0.760	1				
Time-varying characteristics	1.30	0.377	0.243	1			
Worker effect	−0.00	0.502	0.599	−0.476	1		
Estab.-Occup. effect	−0.00	0.397	0.800	0.118	0.333	1	
Residual	0.00	0.196	0.258	−0.000	0.000	0.000	1

Notes: Table reports correlations between components of the decomposition of log hourly earnings into observable characteristics ($X\beta$), unobservable worker heterogeneity (θ), and unobservable establishment-occupation heterogeneity (ψ) according to Equation 5. The estimation sample is the full 100% sample from 2003–2010. The column headers use symbols from the text while row headers provide short definitions.
SOURCE–Authors' calculations based on RAIS microdata.

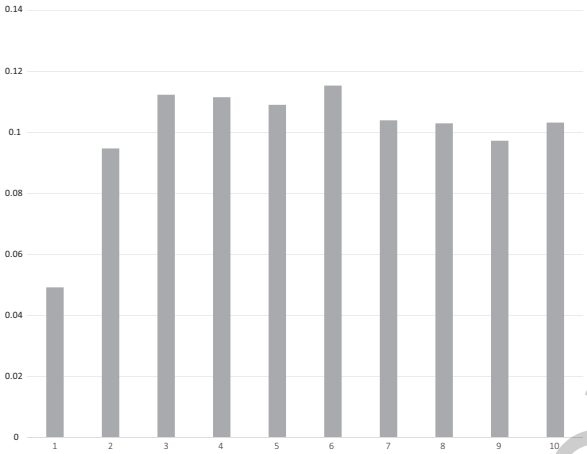


Fig. 1. True Data: Destinations for job transitions from non-employment

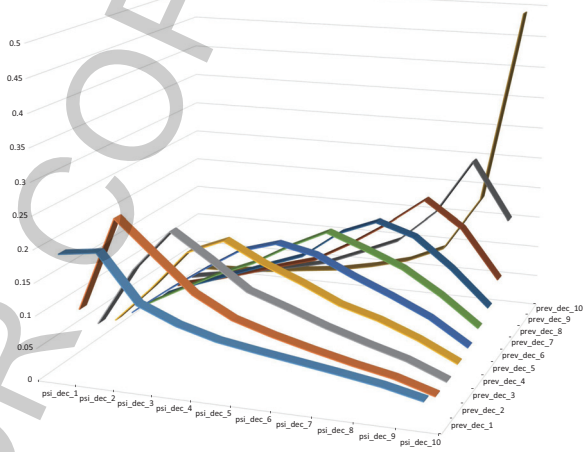


Fig. 2. True Data: Destinations for job-to-job transitions, conditional on origin employer decile. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/SJI-160962)

for which the origin job was with an employer in the first (lowest) decile of the employer effect distribution. These are the lowest-paying jobs. The diagram shows that job-to-job transitions are likely to move workers into jobs in the same decile of the employer effect distribution, or a slightly higher decile. These patterns are very similar to those based on U.S. matched employer-employee data reported in [8].

4. Evaluation

The MWEM mechanism can be used to generate answers to queries from the chosen query set, but the primary output of the algorithm is a synthetic database. By Theorem 2, queries answered using the output synthetic database are guaranteed to have a minimal level of accuracy that depends on the chosen level of privacy protection (ε). Of course, the synthetic data can also be used to answer queries for that were not part of the original query set. Doing so in no way compromises privacy. Differential privacy guarantees are always independent of post-processing.

To evaluate the MWEM algorithm on the RAIS job transition data described in Section 3, I must also specify the query set Q, the number of iterations, T, and the target level of privacy protection, ε. For Q, I chose the set of queries corresponding to all first, second, and third-order marginals. These should guarantee accuracy for queries that would reproduce the patterns in Figs 1 and 2. As discussed below, I am also interested in the distribution of match effects (residuals) within each of these cells, which relies on accuracy with respect to linear queries weighted by the decile cutpoints. However, these queries are not included in Q. The set Q has cardinality $1,306$, which is larger than the cardinality $|D|$. It is not *a priori* clear which set of queries should be used to train K to give the best performance against all possible queries. The optimal number of iterations is very high given these parameters, and I set $T = 300$, which is much lower. In practice, increasing

AUTHOR COPY

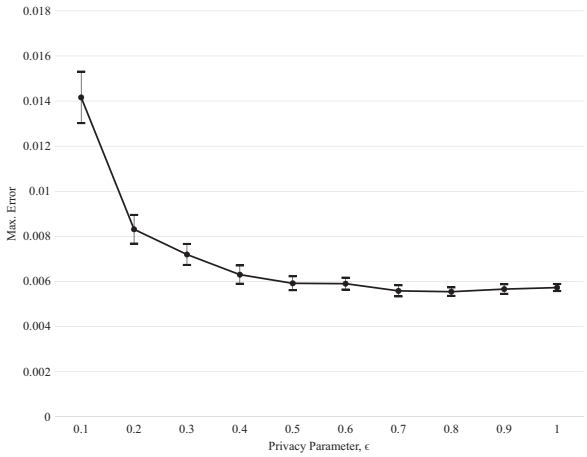


Fig. 3. The maximum error across all queries. The plotted value is the mean of the maximum error across thirty runs of the MWEM algorithm at the given value of the privacy parameter ε. Vertical bars indicate the 95 percent confidence interval around the mean.

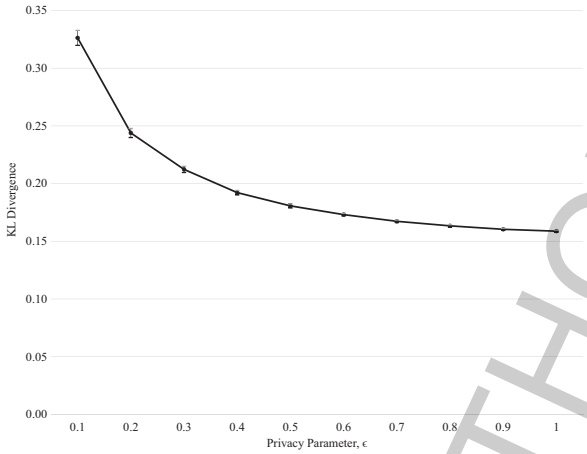


Fig. 4. Mean Kullback-Leibler divergence of synthetic data from the true data. The plotted value is the mean of the maximum error across thirty runs of the MWEM algorithm at the given value of the privacy parameter ε Vertical bars indicate the 95 percent confidence interval around the mean.

T above 300 does not result in an appreciable improvement in performance.

4.1. Maximum error and overall fit

First, I consider how the algorithm performs in terms of the error in answering queries from the chosen query set against the RAIS job transition data, and then consider how much information is lost when using the synthetic data to approximate the true data. Figure 3 displays the average of the maximum absolute error across all queries at each level of ε. The points con-

Fig. 5. Synthetic Data: Destinations for job transitions from non-employment.

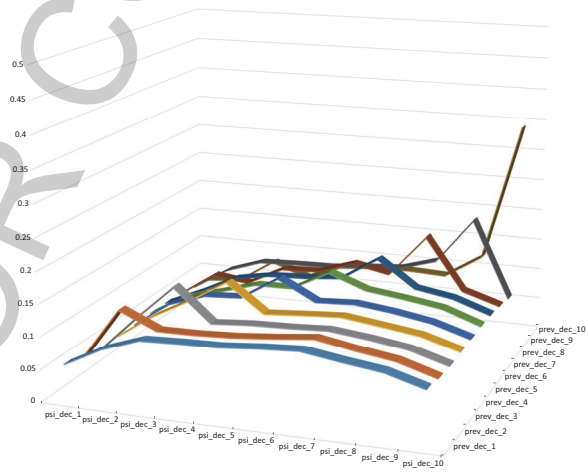


Fig. 6. Synthetic Data: Destinations for job-to-job transitions, conditional on origin employer decile. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/SJI-160962)

nected by the black line represent the average value across thirty runs of the MWEM algorithm at the given level of ε. The vertical bars represent a 95 percent confidence interval around the average.

In the figure, the maximum error is expressed as a fraction of the total database size, n. Except when ε is very low (so privacy protection is very high), the maximum error is slightly above one half of a percent. This is much better than the worst-case guarantee from Theorem 2, though, to be clear, it still indicates a rather poor fit on the worst-case query. The results also show that on this performance metric, there is not much value in increasing ε above 0.4. Beyond that point, reductions in privacy seem to "buy" very little in terms of improved worst-case accuracy.

AUTHOR COPY

Figure 4 plots performance on a different loss metric, the Kullback-Leibler (KL) divergence. The KL-divergence measures the information lost in using the synthetic histogram in place of the true histogram in a manner that does not depend on the particular query set. The KL-divergence is therefore a more general measure of fit. Again, the figure plots the KL-divergence between the synthetic and true data at ten evenly-spaced grid points between $\varepsilon = 0.1$ and $\varepsilon = 1$. The plot shows the average across thirty runs of the MWEM algorithm at the given level of ε. The vertical bars represent a 95 percent confidence interval around the average.

As with the maximal error metric, the KL-divergence is decreasing and convex in privacy loss. However here accuracy continues to improve as privacy increases, albeit with diminishing returns. The overall level of the KL divergence, which is equivalently the expected value of the log-likelihood ratio, is somewhat high.

4.2. *Performance characterizing job-to-job transitions*

Figures 5 and 6 display the synthetic data analogues to the job transition statistics documented in Figs 1 and 2. The synthetic data accurately match the distribution of destinations for transitions from non-employment. For transitions from employment, the synthetic data reflect the tendency for workers to find jobs in the same decile of the employer effect distribution. However, a closer examination of the data shows that the synthetic data are pulled excessively towards a uniform distribution. For this application, the synthetic data convey the basic pattern in the job transition data, but miss some important subtleties in these second order marginals.

4.3. *Performance characterizing expected match effects*

I next consider how well the synthetic data reproduce the distribution of match effects within each type of job transition. These statistics are meaningful, because they shed light on the validity of the identifying assumption of the empirical model of wage determination in Eq. (5). The common version of that assumption is that mobility across employers is exogenous to the wage residual. An implication is that the distribution of wage residuals should not vary with the employer effect on the origin job. [5] use this implication to design tests of the exogenous mobility assumption.

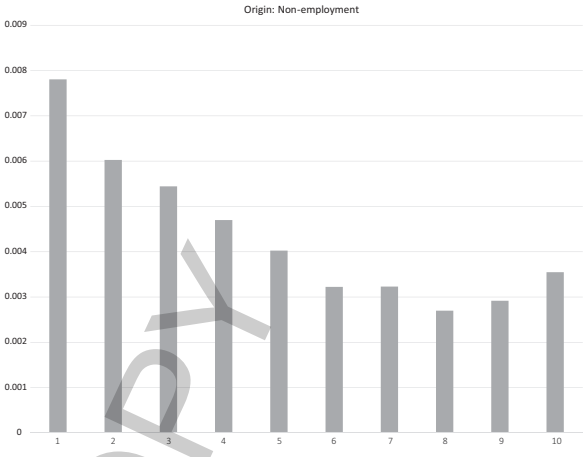


Fig. 7. True Data: Average residual by decile of destination job for transitions from non-employment.

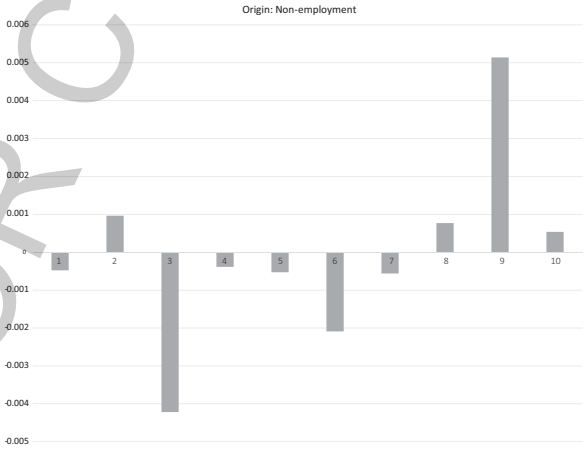
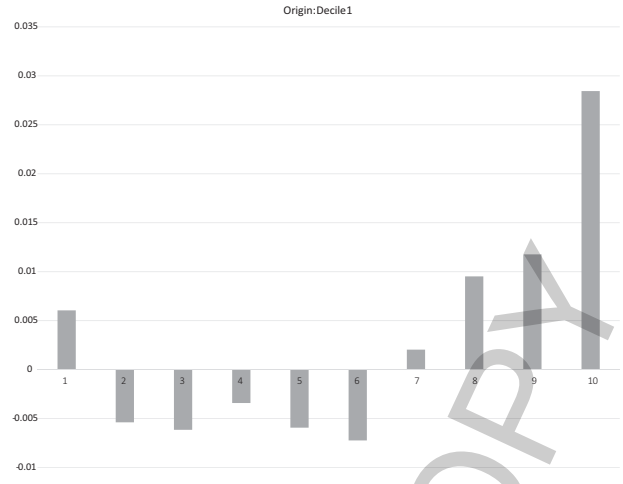


Fig. 8. Synthetic Data: Average residual by decile of destination job for transitions from non-employment.

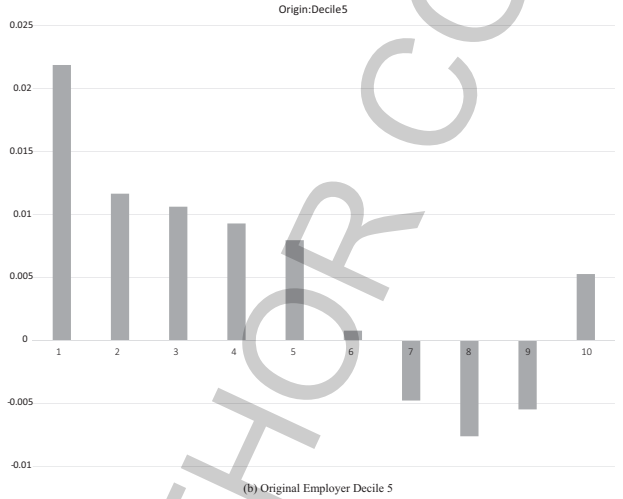
Figures 7 and 5 show, for the true data and synthetic data respectively, the average residual by decile of the destination employer effect for transitions from non-employment. To compute the average wage from the discretized data within each origin-destination employer effect decile combination, I weighted the share of observations in that cell at each decile of the match effect distribution by the midpoint of the match effect within the cell. These are effectively weighted second order marginals. These weighted queries were not included in the query set Q.

In the true data, the average residual is declining with the employer effect. While many factors could explain this pattern, it is consistent with workers being more willing to take jobs with lower-paying firms if they are receiving a higher unexplained portion of pay.

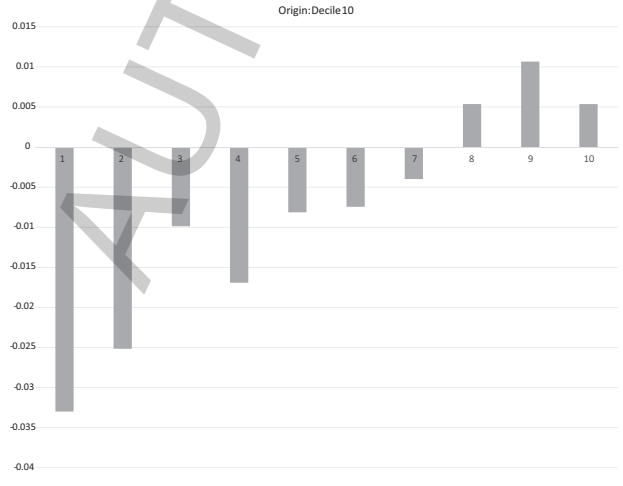
AUTHOR COPY



(a) Original Employer Decile 1



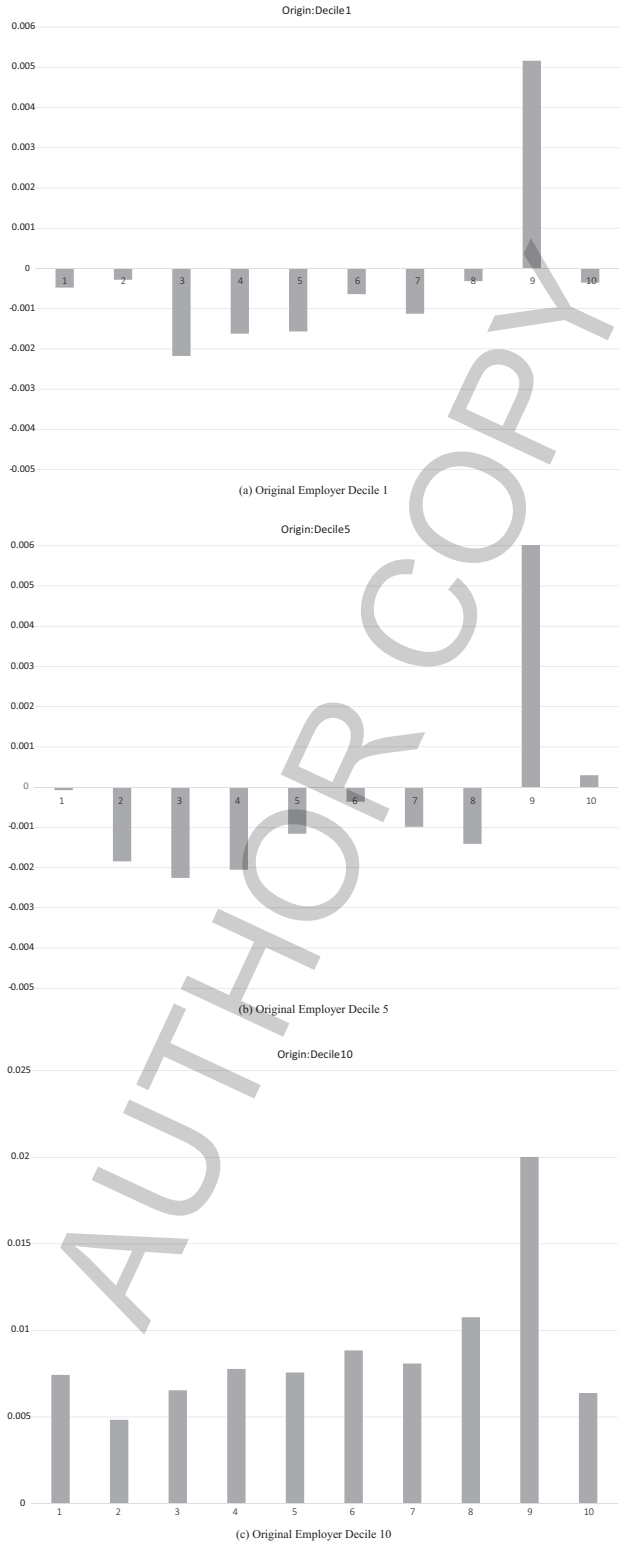
(b) Original Employer Decile 5



(c) Original Employer Decile 10

Fig. 9. True Data: Average residual by decile of destination job for job-to-job transitions.

AUTHOR COPY



AUTHOR COPY

Fig. 10. Synthetic Data: Average residual by decile of destination job for job-to-job transitions.

The synthetic data fare extremely poorly in replicating the overall level of the residuals and also fail to reproduce the strong pattern across deciles.

Figures 9 and 10 report the same statistics, but for transitions from the first, fifth, and tenth decile of the employer effect distribution. The results from the true data in Fig. 9 suggest that, like in the U.S., the data do not support the hypothesis of exogenous mobility. If exogenous mobility holds, then the pattern of wage residuals should be the same across each of the panels in the figure. The plots show that knowledge of the past employer effect is predictive of the wage residual, which is not consistent with exogenous mobility.

One might not draw the same conclusion in an analysis based on the evidence from the synthetic data in Fig. 10. Here, the pattern of residuals across destination employer effect decile is fairly consistent across the panels. A comparison with the true data shows that the patterns in the residuals are inaccurate for all subplots. Therefore, the synthetic data would lead to the wrong impression about the nature of residual variation in the data, and incorrect inference about the model's identifying assumptions.

5. Conclusion

Differentially private data publication has the potential to make confidential data available for public use in a manner that provides strong and verifiable privacy guarantees. Whether the data thus protected are sufficiently useful is a key research question. Furthermore, the range of applications to which differential privacy can be applied is currently limited, but the research frontier is very active.

Related to the present application to the analysis of establishment level data, there are several promising avenues to pursue. First, there is some debate as to the assumptions on the underlying data generating process that must be maintained for the privacy guarantees of differential privacy to hold. In this paper, I have assumed the data are *iid* draws from a general distribution. However, the microdata have a panel dimension. Also, in the linked employer-employee data, there are dependencies across observations that belong to the same worker or to the same establishment. These dependencies mean the guarantees of differential privacy may not hold for a given worker or establishment whose information is in the database [11].

More optimistically, if their methods can be scaled, [12] show that the PMW and Kannan/Frieze algorithms can be used to answer an arbitrary number of graph cut queries. They also show the same method can be used to generate synthetic data that preserve with high probability the cut structure of an underlying graph. This method could address a long-standing obstacle to the production of synthetic matched employer-employee data – that it is not clear how to release data that preserve the relational structure of the data, but that also protect confidentiality of the employers and workers in the data.

This paper sketches only one possible approach to differentially private publication of establishment characteristics. Another approach is to sample synthetic workers, firms, and employment histories using a sequence of differentially private mechanisms, as in [13]. Composibility means differential privacy carries over to the released data. In the context of synthetic call records, [13] show this approach to be quite useful in generating data that can be used in downstream applications that need detailed information on human mobility patterns for simulation-based estimation.

AUTHOR COPY

Acknowledgement

The research in this paper was supported by Alfred P. Sloan Foundation Grant Number G-2015-13903.

References

[1] C. Dwork, Differential Privacy, In: Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP); 2006. pp. 1–12.

[2] C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy. now publishers, Inc.; 2014. Also published as "Foundations and Trends in Theoretical Computer Science" Vol. **9**, Nos. 3–4 (2014) 211–407.

[3] J.M. Abowd and I.M. Schmutte, Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods, Labor Dynamics Institute, Cornell University; 2014, 22.

[4] M. Hardt, K. Ligett and F. Mcsherry, A Simple and Practical Algorithm for Differentially Private Data Release, in: Advances in Neural Information Processing Systems 25, F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, eds, Curran Associates, Inc.; 2012, pp. 2339–2347.

[5] J.M. Abowd, K.L. McKinney and I.M. Schmutte, Modeling Endogenous Mobility in Wage Determination, Cornell University Labor Dynamics Institute Working Paper 23. 2013.

[6] J.M. Abowd, F. Kramarz and D.N. Margolis, High Wage Workers and High Wage Firms, *Econometrica* **67**(2) (1999), 251–333.

[7] D. Card, J. Heining and P. Kline, Workplace Heterogeneity and the Rise of West German Wage Inequality, *The Quarterly Journal of Economics* **128**(3) (2013), 967–1015.

[8] I.M. Schmutte, Job Referral Networks and the Determination of Earnings in Local Labor Markets, *Journal of Labor Economics* **33**(1) (2015), 1–32.

[9] H.R. Hyatt, E. McEntarfer, K. McKinney, S. Tibbets and D. Walton, Job-to-Job (J2J) Flows: New Labor Market Statistics from Linked Employer-Employee Data, US Census Bureau Center for Economic Studies Paper CES-WP-14-34. 2014.

[10] J.M. Abowd, R.H. Creecy and F. Kramarz, Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data, LEHD, U.S. Census Bureau; 2002, TP-2002-06.

[11] D. Kifer and A. Machanavajjhala, No Free Lunch in Data Privacy, In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. SIGMOD '11. New York, NY, USA: ACM; 2011, pp. 193–204.

[12] A. Gupta, A. Roth and J. Ullman, Iterative Constructions and Private Data Release, in: Proceedings of the 9th International Conference on Theory of Cryptography. TCC'12. Berlin, Heidelberg: Springer-Verlag; 2012. pp. 339–356.

[13] D.J. Mir, S. Isaacman, R. Caceres, M. Martonosi and R.N. Wright, DP-WHERE: Differentially private modeling of human mobility, in: Big Data, 2013 IEEE International Conference on; 2013. pp. 580–588.

AUTHOR COPY