

WP. 12
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (ii): Synthetic and hybrid data

**PARAMETER EXPLORATION FOR SYNTHETIC DATA WITH PRIVACY
GUARANTEES FOR *OnTheMap***

Invited Paper

Prepared by John M. Abowd, Johannes Gehrke and Lars Vilhuber, Cornell University

Parameter Exploration for Synthetic Data with Privacy Guarantees for *OnTheMap*

John M. Abowd,^{*} Johannes Gehrke^{**} and Lars Vilhuber^{*}

^{*}School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14853, USA, {john.abowd, lars.vilhuber}@cornell.edu

^{**}Department of Computer Science, Cornell University, Ithaca, NY 14853, USA, johannes@cs.cornell.edu

Abstract: The Census Bureau’s *OnTheMap* application presents detailed geospatial data on worker residences and workplaces. In recent work we developed the algorithms (Machanavajjhala *et al.* 2008), used in the production release of *OnTheMap* Version 3, that achieve probabilistic differential privacy, a formal property for limiting statistical disclosure in the published data. In this paper we explore the parameter space of our algorithm and the resulting analytical validity.

1 Introduction

The United States Census Bureau has published *OnTheMap*, a graphical data application that shows where individuals live and work, since February 2005.¹ The current version (V3) was released in September 2008, and the next version (V4) is scheduled for release in December 2009. The application manipulates a residence/workplace matrix that is resolved to the Census Block level, which permits the definition of custom geographical areas that can then be compared across time and space. Figure 1 provides an example vignette that shows 2006 data for high-earning workers (more than \$40,000/year of individual income in the primary job) from a geographically small area (the Village of Cayuga Heights, near Cornell University). The thermal plot shows the density per square mile of the places of work for the selected individuals who live within the yellow/red boundary of the village. The varying dots show the locations of workplaces according to the scale on the right side of the figure. One of the unique features of these data is that the underlying residence/workplace matrix of job holders has been protected against disclosure of the confidential micro-data using a synthetic data system that offers a formal proof of its privacy protection properties.

¹ The entry page for *OnTheMap* is at <http://lehdm3.did.census.gov/>.

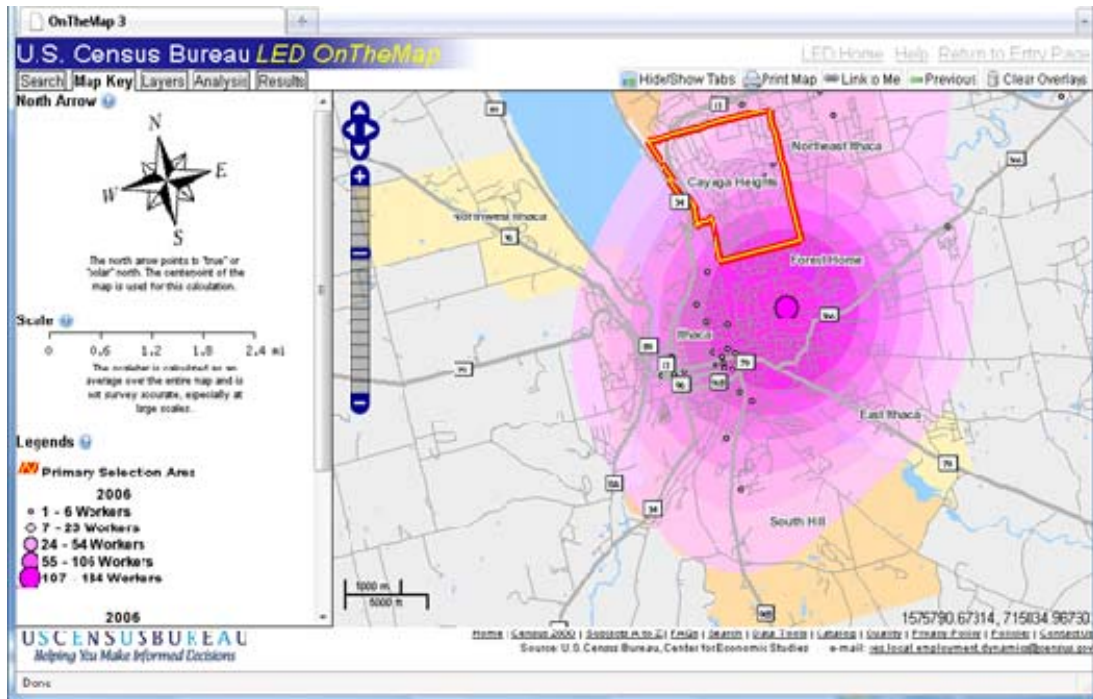


Figure 1: Sample display from *OnTheMap*, Version 3

The formal privacy protection system was developed by Ashwin Machanavajjhala, Daniel Kifer and the authors of this paper (Machanavajjhala *et al.* 2008) specifically for the *OnTheMap* application. Formal privacy protection methods are based on open algorithms with provable properties. They originate from a proof of the impossibility of attaining perfect inferential disclosure protection in the sense of Dalenius (1977). For a more complete discussion in the context of privacy-preserving data mining, see Evfimievski *et al.* (2002), and in the context of differential privacy see Dwork (2006).

This paper provides a brief overview of the formal privacy model used in OTM: probabilistic differential privacy and an empirical analysis of the sensitivity of the data quality to the method's parameters: the privacy-protection limit, the failure probability, and a tuning parameter. All three parameters affect both the protection and data quality. We explore those trade-offs by applying the methods to the public-use data from OTM Version 3. The application reveals, as expected, that the quality of the synthetic data is very sensitive to the strictness of the differential privacy standard applied. Somewhat surprisingly, the data quality is less affected by the tuning parameter.

2 Synthetic Data, Randomized Sanitizers and Probabilistic Differential Privacy

Synthetic data (Rubin 1993; Little 1993) are created by estimating the posterior predictive distribution (PPD) of the release data given the confidential data; then sampling release data from the PPD conditioning on the actual confidential values. The PPD is a parameter-free forecasting model for new values of the complete data matrix that conditions on all values of the underlying confidential data. In general, the properties of synthetic data can be assessed using the transition matrix $Pr[\tilde{X}|X]$ where \tilde{X} are the synthetic samples and X are the confidential data.

A randomized sanitizer creates a conditional probability distribution for the release data given the confidential data, $Pr[\tilde{X}|X]$, from the function $San[X, U]: (X, U) \rightarrow \tilde{X}$ which maps the confidential data X and random noise from a pre-specified distribution, U , into the release data \tilde{X} . The randomness in a sanitizer is induced by the properties of the distribution of U . The PPD is just a particular randomized sanitizer.

2.1 How Synthetic Data Leak Information

Suppose $x^{(1)}$ and $x^{(2)}$ are two confidential data sets that differ in a single element—observation and value of a variable. Then, $Pr[\tilde{X}|x^{(1)}] \neq Pr[\tilde{X}|x^{(2)}]$ for any randomized sanitizer or synthesizer. If the exact confidential data are released, then $Pr[\tilde{X}|X] = I$, where I is the identity matrix. Every element of the confidential data maps to one, and only one, element of the released data. Any statistical disclosure limitation (SDL) procedure that changes the input data in order to produce the release data can be represented with a transition matrix. In SDL, this formalization of the protection process was called post-randomization by Kooiman et al (1997); however, the usefulness of this construct goes beyond their initial formalization.

Dwork (2006) and Chawla *et al.* (2005) showed that Dalenius’s original (1977) notion of an inferential disclosure could be formalized in terms of this same transition matrix. In particular, consider the equation:

$$\frac{\frac{Pr[X = x^{(1)}|\tilde{X} = \tilde{x}]}{Pr[X = x^{(2)}|\tilde{X} = \tilde{x}]}}{\frac{Pr[X = x^{(1)}]}{Pr[X = x^{(2)]}}} = \frac{Pr[\tilde{X} = \tilde{x}|X = x^{(1)}]}{Pr[\tilde{X} = \tilde{x}|X = x^{(2)}]}$$

for some realization of the synthesizer, \tilde{x} , and any two realizations of the confidential data $x^{(1)}$ and $x^{(2)}$. The left-hand side of this equation is the posterior odds ratio for the gain in information about the difference between $x^{(1)}$ and $x^{(2)}$, given the release \tilde{x} . The right-hand side of this equation is the quantity used to define privacy in cryptography-based SDL methods. Dalenius called the left-hand side an inferential

disclosure. Privacy-preserving datamining calls the right-hand side the privacy limit, for specific definitions of $x^{(1)}$ and $x^{(2)}$.

In ε -differential privacy systems, $x^{(1)}$ and $x^{(2)}$ differ by a single element, or row depending upon the application, $\ln \left[\frac{Pr[\tilde{X}=\tilde{x}|X=x^{(1)}]}{Pr[\tilde{X}=\tilde{x}|X=x^{(2)}]} \right]$ is bounded by ε . Hence, ε -differential privacy bounds the inferential disclosure by bounding the log posterior odds for the inference about a particular data element for a particular observation, given all the remaining confidential data—an extraordinarily comprehensive information set for the attacker/user. In probabilistic differential privacy (Machanavajjhala et al. 2008), $\ln \left[\frac{Pr[\tilde{X}=\tilde{x}|X=x^{(1)}]}{Pr[\tilde{X}=\tilde{x}|X=x^{(2)}]} \right]$ is bounded by ε with probability $1 - \delta$ for the same assumptions on $x^{(1)}$ and $x^{(2)}$. The extension to ε -privacy

(Machanavajjhala et al. 2009) bounds $\frac{Pr[X=x^{(1)}|\tilde{X}=\tilde{x}]}{Pr[X=x^{(2)}|\tilde{X}=\tilde{x}]}$ directly for increasingly strenuous

definitions of the prior information used to construct the denominator. The most strenuous of those definitions duplicates the ε -differential privacy bound.

2.2 Formal Properties of Probabilistic Differential Privacy

We begin by defining ε -Differential Privacy:

Definition (ε - Differential Privacy): Let \mathbf{A} be a randomized algorithm, let \mathbf{S} be the set of all possible outputs of the algorithm, and let $\varepsilon > 0$. The algorithm \mathbf{A} satisfies ε - differential privacy if for all pairs of data sets (D_1, D_2) that differ in exactly one row,

$$\forall S \in \mathbf{S}, \frac{P(\mathbf{A}(D_1)) = S}{P(\mathbf{A}(D_2)) = S} \leq e^\varepsilon \text{ or } \ln \left| \frac{P(\mathbf{A}(D_1)) = S}{P(\mathbf{A}(D_2)) = S} \right| < \varepsilon.$$

Differential privacy (Dwork, 2006, and Chawla, et al. 2005) is difficult to maintain in sparse applications, for example, when geographically nearby blocks have very different posterior probabilities. To formalize this idea, Machanavajjhala et al (2008), developed the concept of a Disclosure Set, which describes the conditions under which ε -Differential Privacy fails. Specifically:

Definition (Disclosure Set) : Let D be a table and \mathbf{D} be the set of tables that differ from D in at most one row. Let \mathbf{A} be a randomized algorithm and \mathbf{S} be the space of outputs of the algorithm \mathbf{A} . The disclosure set of D , denoted $\text{Disc}(D, \varepsilon)$, is

$$\left\{ S \in \mathbf{S} \mid \exists X_1, X_2 \in \mathbf{D}(D), |X_1 \setminus X_2| = 1 \wedge \left| \ln \frac{P(\mathbf{A}(X_1) = S)}{P(\mathbf{A}(X_2) = S)} \right| > \varepsilon \right\}.$$

When differential privacy fails in situations where the underlying table is very sparse, it is usually the case that the failure comes from disclosure sets that have a very small probability of occurring in the released data. Probabilistic Differential Privacy (PDP) formalizes this idea:

Definition (Probabilistic Differential Privacy) : Let \mathbf{A} be a randomized algorithm and \mathbf{S} be the space of outputs of \mathbf{A} . Let $\varepsilon > 0$ and $0 < \delta < 1$ be constants. Then \mathbf{A} satisfies (ε, δ) -probabilistic differential privacy (or (ε, δ) -pdp) if for all tables D ,

$$P(\mathbf{A}(D) \in \text{Disc}(D, \varepsilon)) \leq \delta.$$

PDP allows one to control the probability that differential privacy fails. The analytical validity of sparse applications can be controlled with PDP because the restrictions on the prior used in the synthesizer are reasonable for use with sparse tables. This is the idea that we will illustrate with OTM below.

2.3 The OTM Synthesizer

The synthetic OTM data are based on the Multinomial-Dirichlet posterior predictive distribution defined for each workplace block j :

$$\begin{aligned} \mathbf{n} &= (n_1, \dots, n_k), n = \sum n_i \\ \boldsymbol{\alpha} &= (\alpha_1, \dots, \alpha_k), \alpha_0 = \sum \alpha_i \\ \boldsymbol{\pi} &= (\pi_1, \dots, \pi_k) \\ \mathbf{n} &\sim \text{M}(n, \boldsymbol{\pi}) \\ \boldsymbol{\pi} &\sim \text{D}(\boldsymbol{\alpha}), \text{ a priori} \\ \boldsymbol{\pi} &\sim \text{D}(\boldsymbol{\alpha} + \mathbf{n}), \text{ a posteriori} \\ \mathbf{m} &= (m_1, \dots, m_k), m = \sum m_i \\ \mathbf{m} &\sim \text{M}(m, \boldsymbol{\pi}) \end{aligned}$$

The vector \mathbf{n} is the actual counts of individuals living in block i whose place of work is block j , where the workplace block subscripts have been suppressed. The prior sample sizes for the Dirichlet are the vector $\boldsymbol{\alpha}$. In OTM, the workplace employment

counts come directly from the confidentiality-protected Quarterly Workforce Indicators. For a given workplace block, m is the total employment count for all employers on the block, which is not synthesized, but has been protected by the QWI noise-infusion system (Abowd *et al.* 2009), and the vector \mathbf{m} contains the synthetic residence block distribution.

To apply PDP to the OTM synthesizer, the parameters ε and δ are specified, then an α vector is computed that conforms to these parameters. Hence, the confidentiality protections are embedded in the Dirichlet prior for the Multinomial probabilities. The tuning parameter p_{min} is used to reduce the dimensionality of the domain of the synthesizer by randomly retaining points where $n_i = 0$ with probability at least p_{min} . Pruning the synthesizer’s domain has a privacy penalty. The differential privacy parameter ε must be adjusted to $\varepsilon' = \varepsilon + \ln\left(\frac{1}{p_{min}}\right) + \ln(\max[\alpha_i]) \ln 2$.

2.3.1 Consequences of varying the parameters of PDP

In this paper we use the OTM synthesizer applied to tract-to-tract residence and workplace locations to measure the consequences of varying ε and p_{min} .² The protection afforded by PDP is not affected by the coarsening from blocks to tracts; however, using tracts permits us to evaluate the analytical validity without further assumptions by computing the integrated mean squared error for each workplace block, then averaging those IMSEs over all workplaces using the employment total in the tract as the averaging weight.

3 Data Sources and Definitions

The basic micro-data sources for OTM are described in Abowd et al (2009) and Wu and Graham (2009). Workplace and residence geographies are defined using Census blocks. Statistical analysis to estimate the PPD is based on Census tract-to-tract relations. There are 8.2 million blocks and 65,000 tracts in the U.S. Every workplace block with positive employment has its own synthesizer. We used the OTM public use data for the state of Minnesota, as in Machanavajjhala et al. (2008), except that we used the version 3 data.³

4 Results

Figure 2 shows the relation between the differential privacy parameter, ε , on the horizontal axis and the square root of the average IMSE for tracts with different employment levels on the vertical axis, holding $\delta = 0.000001$. Lower values of the

² The parameter d was held constant to allow us to study the trade-off between the other two parameters, which are tightly linked in the PDP model.

³ These data are available for download at <http://www.vrdc.cornell.edu/onthemap/doc/>.

root average IMSE imply better analytical validity. The value on the vertical axis may be interpreted as one-half the average width of the one-standard-error confidence band surrounding an estimate of the proportion of individuals living a particular census tract who work in a destination tract with the indicated employment level. As can be seen from the figure, setting the log-odds limit at 2 ($\epsilon = 2$), involves a very substantial average data quality penalty as compared to setting the log odds limit at 4 or 4.6 (posterior odds of 100:1). Loosening this parameter further, has no appreciable effect on the data quality.

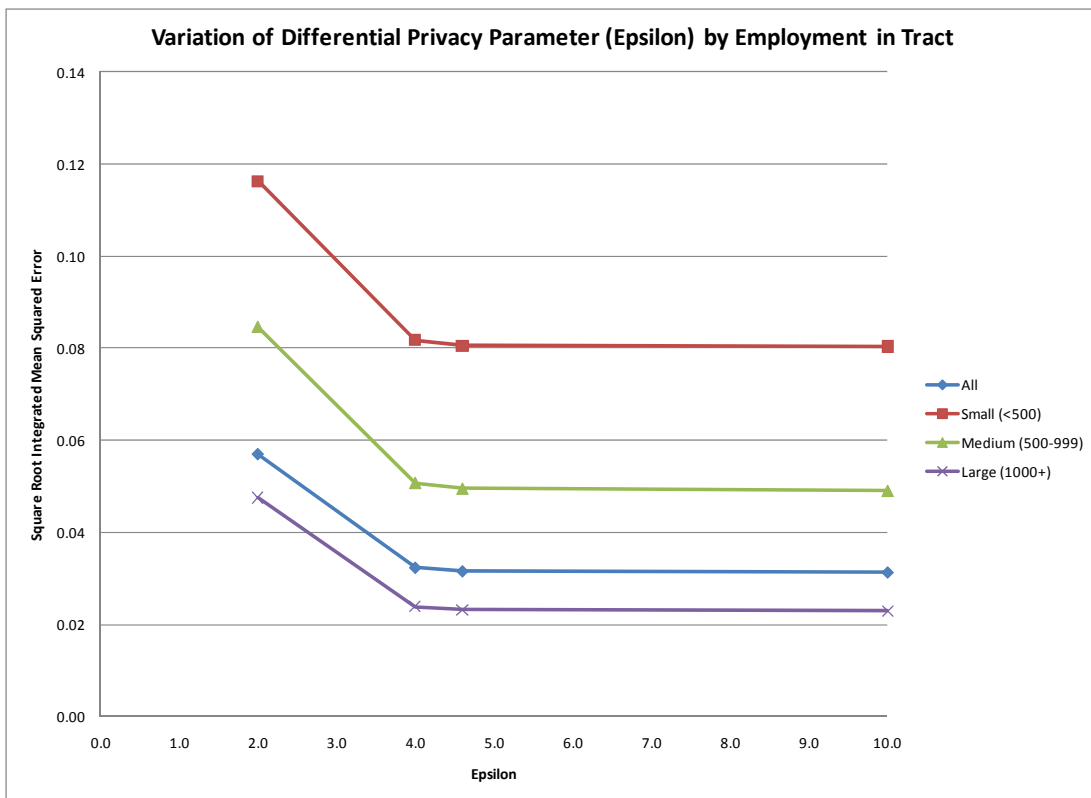


Figure 2

Figure 3 shows the effects of varying the minimum retention probability, which is plotted on the horizontal axis, on the square root of the average IMSE. We only investigated the range of $0.01 \leq p_{min} \leq 0.1$ in this paper since earlier work suggested that larger values would be very costly. The figure shows that this parameter has relatively little effect on the analytical validity in the range of practical values. The root average IMSE rises very slightly for each employment size class as the minimum retention probability is increased. This result is surprising and warrants further investigation. Other analytical validity measures, such as the Kullback-Leibler divergence, which cannot be calculated reliably for tract-to-tract assessment because there are too many sampling zeroes, have been sensitive to the level of p_{min} within this range.

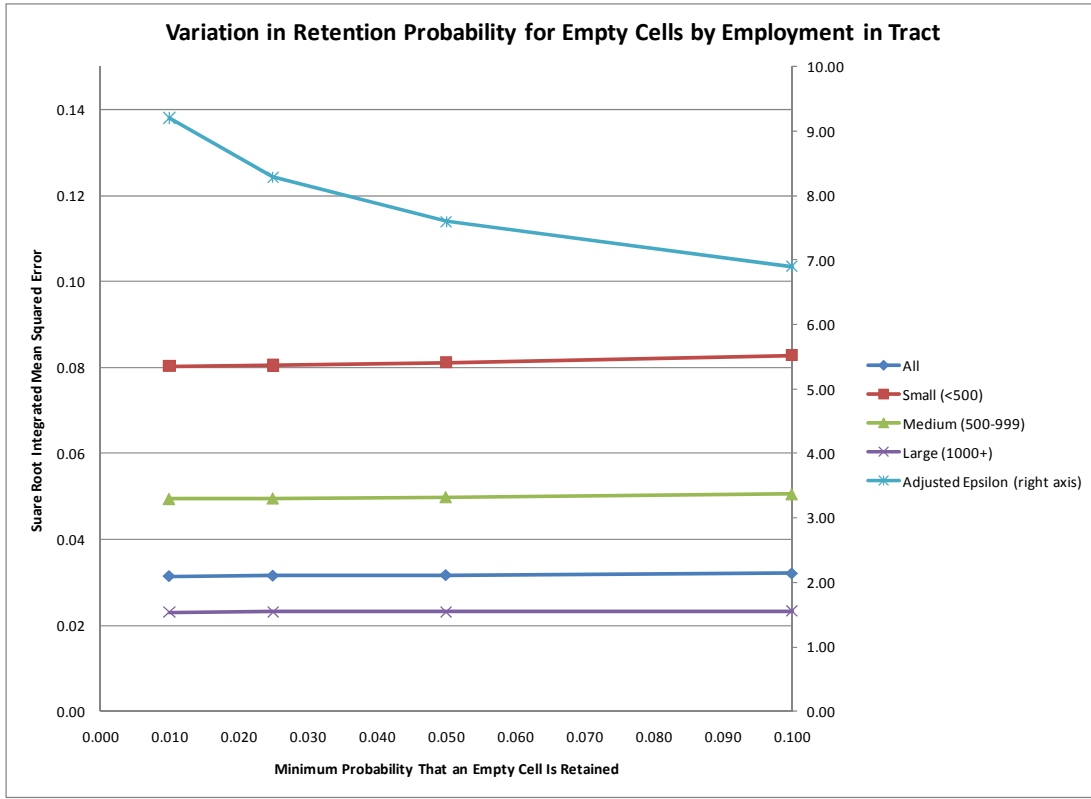


Figure 3

5 Conclusion

We have conducted an empirical evaluation of the sensitivity of the data quality to the level of privacy protection as defined by Probabilistic Differential Privacy. The empirical analysis applies to the public-use data from the U.S. Census Bureau's *OnTheMap* application, which have the same structure as the underlying confidential data and can, therefore, be used as the basis for an evaluation of the methodology used to protect them. Measuring analytical validity by the square root of the average integrated mean squared error for a tract-to-tract table of residence and workplace locations, we find that the validity is very sensitive to the protection level for small values of the log odds limit, but at intermediate values, above 4, the data quality stops deteriorating. The PDP tuning parameter, the minimum probability of retaining an empty cell in the domain of the posterior, did not affect the data quality in these experiments. The latter result is surprising and warrants further study.

Acknowledgements

This research uses data from the United States Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following grants: National Science Foundation SES-9978093, SES-0339191, and ITR-0427889; National Institute on Aging AG018854; and grants from the Alfred P. Sloan Foundation. The authors also acknowledge partial direct support by NSF grants CNS-0627680, SES-0820349, SES-0922005, and SES-0922494, and by the Census Bureau. No confidential data were used in this paper. All public-use *OnTheMap* data can be accessed at <http://lehdm3.did.census.gov/themap3/> or downloaded from the Cornell VirtualRDC at <http://www.vrdc.cornell.edu/onthemap/doc>.

References

- 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.
- Abowd, J. and L. Vilhuber (2008) "How Protective are Synthetic Data," in J. Domingo-Ferrer and Y. Saygun, eds., *Privacy in Statistical Databases 2008* (Berlin: Springer-Verlag), pp. 239-246.
- Chawla, S., C. Dwork F. McSherry, A. Smith, and H. Wee (2005) "Towards privacy in public databases," in *Proceedings of the 2nd Theory of Cryptography Conference*.
- Evfimievski, A. , R. Srikant, R. Agrawal, and J. Gehrke (2002) "Privacy Preserving Mining of Association Rules," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2002).
- Dalenius, T. (1997) "Towards a methodology for statistical disclosure control," *Statistik Tidskrift (Statistical Review)*, pp. 429-44.
- Little, R. (1993) "Statistical Analysis of Masked Data," *Journal of Official Statistics*, Vol. 9, pp. 407-26.
- Dwork, C "(2006) Differential Privacy," *33rd International Colloquium on Automata, Languages, and Programming—ICALP 2006, Part II*, 1-12.
- Kooiman, P., Willenborg, L.C.R.J. and Gouweleeuw, J.M., "PRAM: a method for disclosure limitation of microdata," Research paper no. 9705, Statistics Netherlands, (1997).
- Machanavajjhala, A., J. Gehrke, and M. Götz (2009) "Data Publishing against Realistic Adversaries" VDBL 2009.
- Machanavajjhala, A. D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. (2008) "Privacy: From theory to practice on the map," ICDE, 2008, pp. 277-286.

- Rubin, D. (1993) "Discussion Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol. 9, pp. 461-68.
- Wu, Jeremy S. and Graham, Matthew R (2009) "OnTheMap: An Innovative Mapping and Reporting Tool," *The United Nations Statistical Commission Seminar on Innovations in Official Statistics* (February 2009) available online at http://unstats.un.org/unsd/statcom/statcom_09/seminars/innovation/Innovation%20Seminar/USA-OntheMap.pdf (cited November 24, 2009).