1-17-2013

Presentation: Revisiting the Economics of Privacy: Population Statistics and Privacy as Public Goods

John Abowd
Cornell University, John.Abowd@cornell.edu

Presentation: Revisiting the Economics of Privacy: Population Statistics and Privacy as Public Goods

Abstract
Anonymization and data quality are intimately linked. Although this link has been properly acknowledged in the Computer Science and Statistical Disclosure Limitation literatures, economics offers a framework for formalizing the linkage and analyzing optimal decisions and equilibrium outcomes.

Keywords
Data Privacy, Confidentiality Protection, Economics

Comments
Data Linkage and Anonymisation Scoping Meeting, Issac Newton Institute for Mathematical Sciences, Cambridge, UK

Revisiting the Economics of Privacy: Population Statistics and Privacy as Public Goods

John M. Abowd
Cornell University
January 17, 2013

Acknowledgements and Disclaimer

- This research uses data from the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program, which was partially supported by the following grants: National Science Foundation (NSF) SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging AG018854; and grants from the Alfred P. Sloan Foundation
- I also acknowledge partial direct support by NSF grants CNS-0627680, SES-0820349, SES-0922005, SES-0922494, BCS-0941226, SES-1042181, TC-1012593, and SES 1131848; and by the Census Bureau
- All confidential data used for this presentation were reviewed using the Census Bureau's disclosure avoidance protocols
- The opinions expressed in this presentation are those of the author and neither the National Science Foundation nor the Census Bureau

Colleagues and Collaborators

Fredrik Andersson, Matthew Armstrong, *Sasan Bakhtiari*, Patti Becker, *Gary Benedetto*, *Melissa Bjelland*, Chet Bowie, Holly Brown, Evan Buntrock, *Hyowook Chiang*, *Stephen Ciccarella*, Cynthia Clark, Rob Creecy, *Lisa Dragoset*, Chuncui Fan, John Fattaleh, Colleen Flannery, Lucia Foster, *Matthew Freedman*, *Monica Garcia-Perez*, Johannes Gehrke, Nancy Gordon, *Kaj Gittings*, Matthew Graham, Robert Groves, Owen Haaga, Hermann Habermann, John Haltiwanger, Heath Hayward, *Tomeka Hill*, Henry Hyatt, Emily Isenberg, Ron Jarmin, Dan Kifer, C. Louis Kincannon, Shawn Klimek, Fredrick Knickerbocker, Mark Kutzbach, Walter Kydd, Julia Lane, *Paul Lengermann*, Tao Li, Cindy Ma, Ashwin Machanavajjhala, Erika McEntarfer, Kevin McKinney, Thomas Mesenbourg, Jeronimo Mulato, *Nicole Nestoriak*, Camille Norwood, Ron Prevost, Kenneth Prewitt, George Putnam, Kalyani Raghunathan, Uma Radhakrishnan, Arnie Reznek, *Bryan Ricchetti*, Jerry Reiter, Marc Roemer, Kristin Sandusky, *Ian Schmutte*, Matthew Schneider, Rob Sienkiewicz, *Liliana Sousa*, *Bryce Stephens*, *Martha Stinson*, *Michael Strain*, Stephen Tibbets, Lars Vilhuber, J. Preston Waite, Chip Walker, *Doug Webber*, Dan Weinberg, Bill Winkler, *Simon Woodcock*, Jeremy Wu, Laura Zayatz, Chen Zhao, Nellie Zhao, Lingwen Zheng, and Chaoling Zheng

Italics = earned Ph.D. while interning at LEHD

Overview

- Anonymization and data quality are intimately linked
- Although this link has been properly acknowledged in the CS and SDL literatures, economics offers a framework for formalizing the linkage and analyzing optimal decisions and equilibrium outcomes

Technology

$$f(I,\phi;b)=0 \text{ from } \left\{(I,\phi)\middle|\max_{\tilde{\phi}\le\phi} I(\tilde{\phi}) \text{ s.t. } I(0)\le b\right\}$$

where

I is a measure of the information in a data publication

ϕ is a measure of the privacy/confidentiality protection

b is the total resource limit

Technology of Anonymization

- Privacy (CS)/confidentiality (SDL) controls on data publication can be described formally as a production possibility frontier
- A PPF measures the maximum attainable data quality when the privacy controls are parameterized as ϕ, (-ε from the differential privacy viewpoint)
- This is related to risk-utility curves in statistics but the formalization is more demanding

Preferences

$$v_i\left(y_i; I, \phi, p\right) = \max_x u_i\left(x, I, \phi\right) \text{ s.t. } x^T p \leq y_i$$

$$i = 1, \ldots, N$$

where

u_i is consumer i's direct utility function

v_i is consumer i's indirect utility function

y_i is consumer i's income

I, ϕ are the public goods (data information and privacy)

x_i is the chosen private good bundle

p is the vector of private good prices

Public Goods and Private Goods

- My formulation of the problem makes both the data publication (I) and the privacy associated with the publication (ϕ) public goods.
- No privileged access to the data (think: public-use tables or series)
- Equal protection of all consumer/citizens

Samuelson (1954) Equilibrium

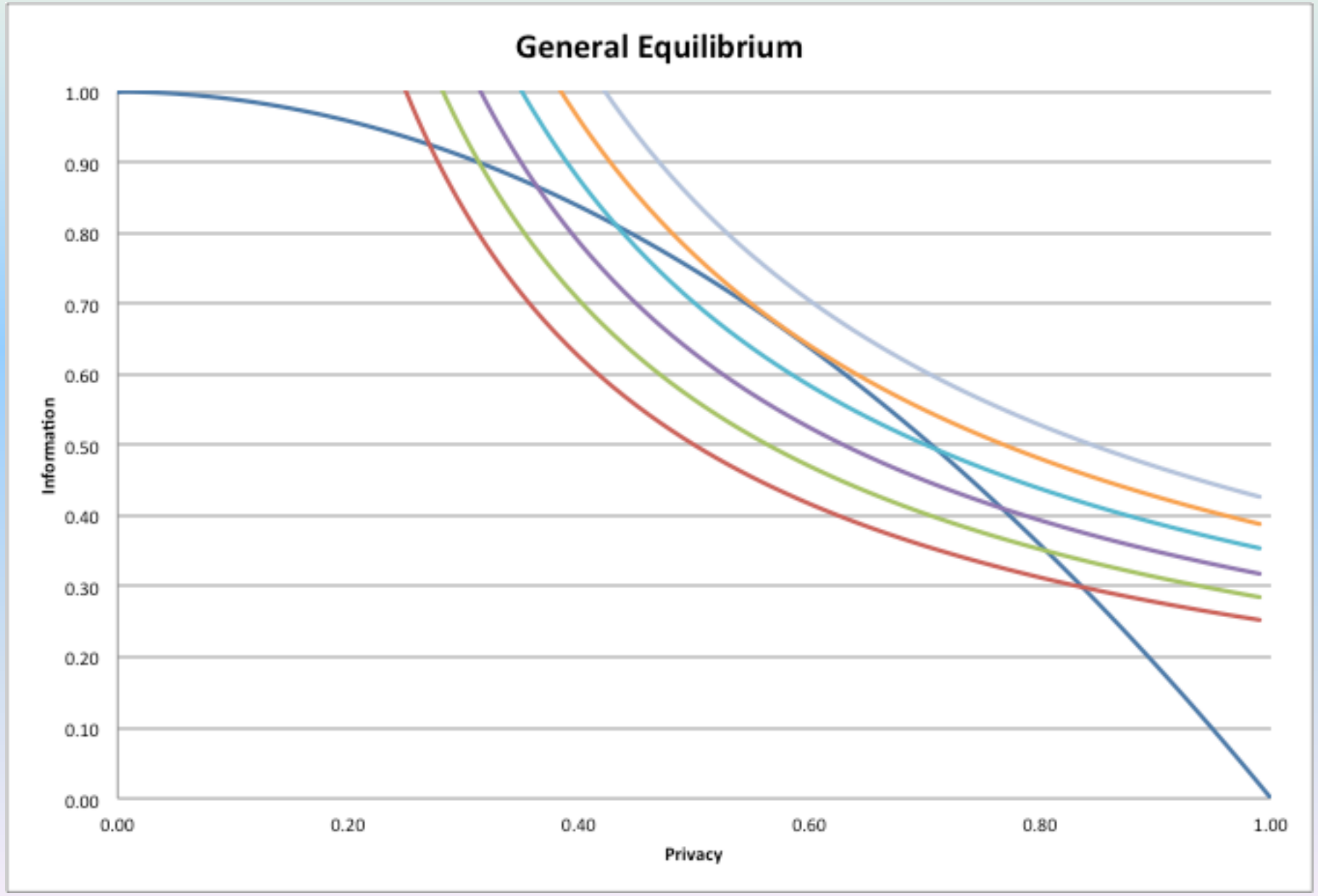
$$SWF: \sum_{i=1}^{n} v_i\left(y_i, I, \phi, p\right)$$

$$PPF: f\left(I, \phi, b\right) = 0$$

Optimal production of the public goods $\left(I^0, \phi^0\right)$

maximize SWF subject to PPF.

$$\frac{\dfrac{\partial f\left(I^0, \phi^0, b\right)}{\partial \phi}}{\dfrac{\partial f\left(I^0, \phi^0, b\right)}{\partial I}} = \frac{\dfrac{\partial}{\partial \phi}\sum_{i=1}^{N} v_i\left(y_i, I^0, \phi^0, p\right)}{\dfrac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i, I^0, \phi^0, p\right)}$$


General Equilibrium
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Information
0.00
0.20
0.40
0.60
0.80
1.00
Privacy

Implications of Public Good Model

- With zero collection costs (PPF depends only on the privacy technology), always conduct a census (or, use all the administrative records)
- Straightforward to relax this, but not helpful
- Set the marginal rate of transformation (slope of the PPF) equal to the ratio of the sums of the marginal utilities of the consumers (not the marginal rate of substitution as with a private good)
- Private provision of *I* fails; it is undersupplied, privacy is oversupplied

Special Case: Separable Utility

$$\frac{\dfrac{\partial f\left(I^0,\phi^0,b\right)}{\partial \phi}}{\dfrac{\partial f\left(I^0,\phi^0,b\right)}{\partial I}} = \frac{\dfrac{\partial}{\partial \phi}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)}{\dfrac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)} = \frac{\sum_{i=1}^{N}\dfrac{\partial v_i}{\partial \phi}\left(I^0,\phi^0,p\right)}{\sum_{i=1}^{N}\dfrac{\partial v_i}{\partial I}\left(I^0,\phi^0,p\right)} = \frac{\overline{\dfrac{\partial v}{\partial \phi}}}{\overline{\dfrac{\partial v}{\partial I}}}$$

- The optimal choice of data information and privacy depends upon the ratio of average marginal utilities.
- Optimal choice caters to the average consumer (not an extreme consumer)

Special Case: Separable, Identical Utility

$$\frac{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial \phi}}{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial I}} = \frac{\frac{\partial}{\partial \phi}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)}{\frac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)} = \frac{\sum_{i=1}^{N}\frac{\partial v}{\partial \phi}\left(I^0,\phi^0,p\right)}{\sum_{i=1}^{N}\frac{\partial v}{\partial I}\left(I^0,\phi^0,p\right)} = \frac{\frac{\partial v}{\partial \phi}\left(I^0,\phi^0,p\right)}{\frac{\partial v}{\partial I}\left(I^0,\phi^0,p\right)}$$

- The optimal choice can be determined by the representative consumer even though all consumers are not identical, so there is still demand for the information

Special Case: Non-separable Quadratic Utility I

$$\frac{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial \phi}}{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial I}} = \frac{\frac{\partial}{\partial \phi}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)}{\frac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i,I^0,\phi^0,p\right)} = \frac{\sum_{i=1}^{N}\delta_i y_i \phi^0}{\sum_{i=1}^{N}\eta_i y_i I^0} = \frac{\overline{y}_\delta \phi^0}{\overline{y}_\eta I^0}$$

- The optimal choice depends on the ratio of weighted means of income, weighted by privacy preferences in the numerator and by information preferences in the denominator

Special Case: Non-separable Quadratic Utility II

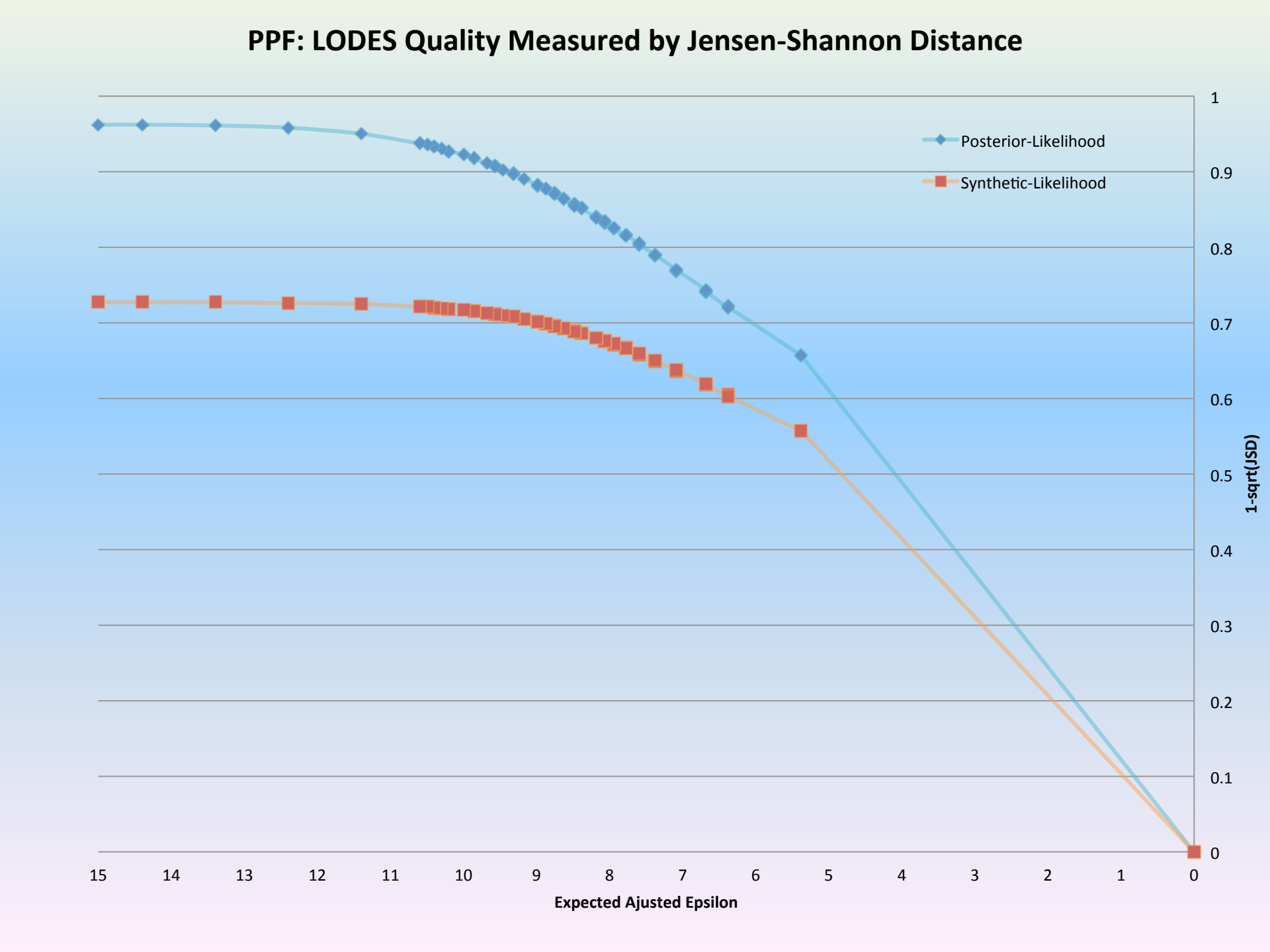
$$\frac{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial \phi}}{\frac{\partial f\left(I^0,\phi^0,b\right)}{\partial I}} = \frac{\frac{\partial}{\partial \phi}\sum_{i=1}^{N} v_i\left(y_i-\bar{y},I^0,\phi^0\right)}{\frac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i-\bar{y},I^0,\phi^0\right)} = \frac{\sum_{i=1}^{N}\left[\delta_i\left(y_i-\bar{y}\right)\phi^0\right]}{\sum_{i=1}^{N}\left[\eta_i\left(y_i-\bar{y}\right)I^0\right]} = \frac{\mathrm{Cov}\left[\delta_i,y_i\right]\phi^0}{\mathrm{Cov}\left[\eta_i,y_i\right]I^0}$$

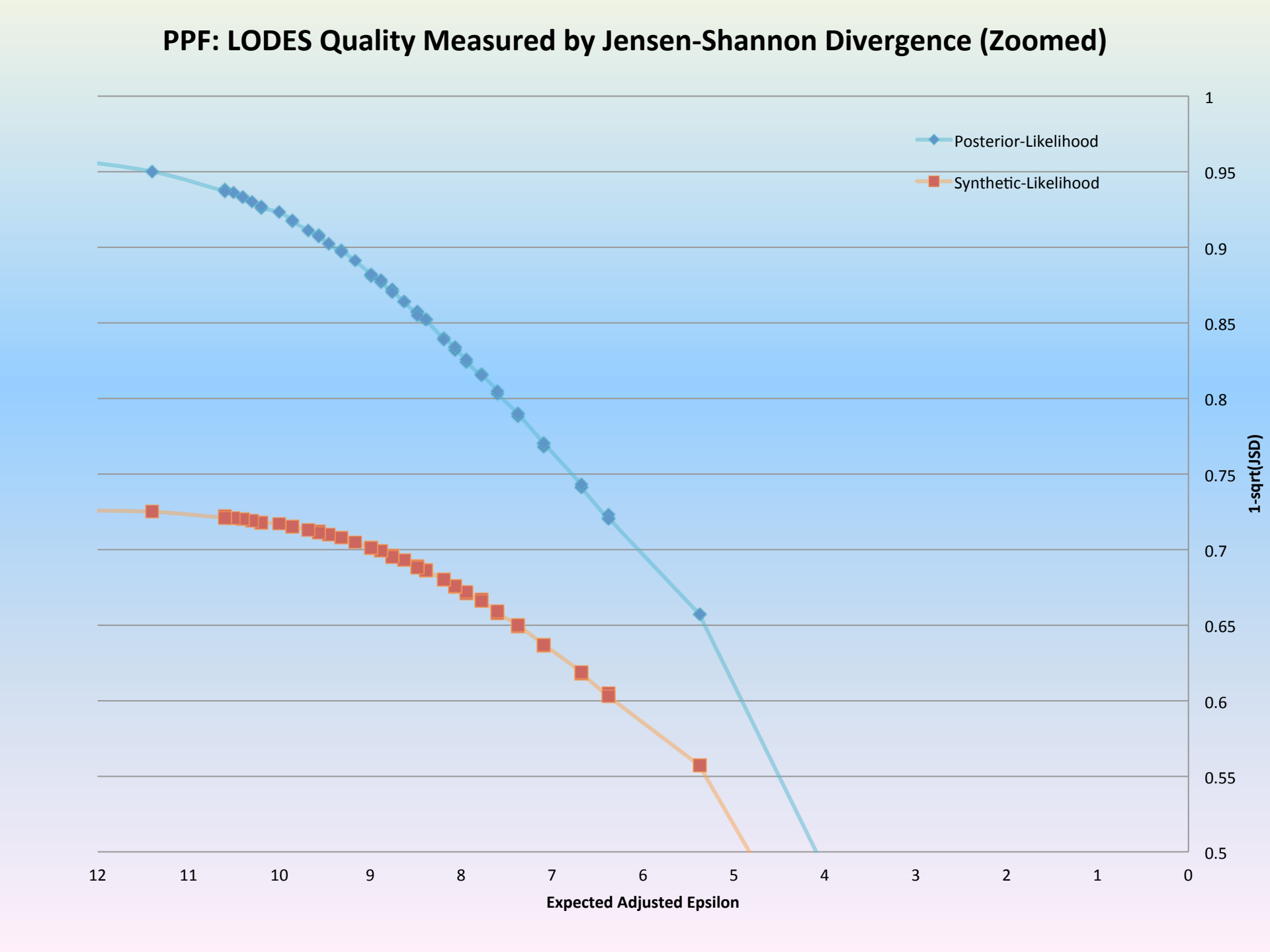
- The optimal choice depends on the ratio of covariances of preferences towards privacy (numerator) and information (denominator) with income

Example 1 PPF

$$I_{JSD}(\phi) = 1 - \sqrt{0.5\sum_k \pi_k \log_2 \frac{\pi_k}{0.5\hat{\pi}_k(\phi) + 0.5\pi_k} + 0.5\sum_k \hat{\pi}_k(\phi) \log_2 \frac{\hat{\pi}_k(\phi)}{0.5\hat{\pi}_k(\phi) + 0.5\pi_k}}$$

- Based on the Jensen-Shannon distance between the true probabilities over a grid k = 1, …, K (π_k) and the probability in each cell after protection $(\hat{\pi}_k(\phi))$
- Note that the total information from a census of *N* individuals is normalized to 1, this would change if the size of the population changes, general form is b(*N*)


PPF: LODES Quality Measured by Jensen-Shannon Distance
Posterior-Likelihood
Synthetic-Likelihood
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1-sqrt(JSD)
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Expected Ajusted Epsilon

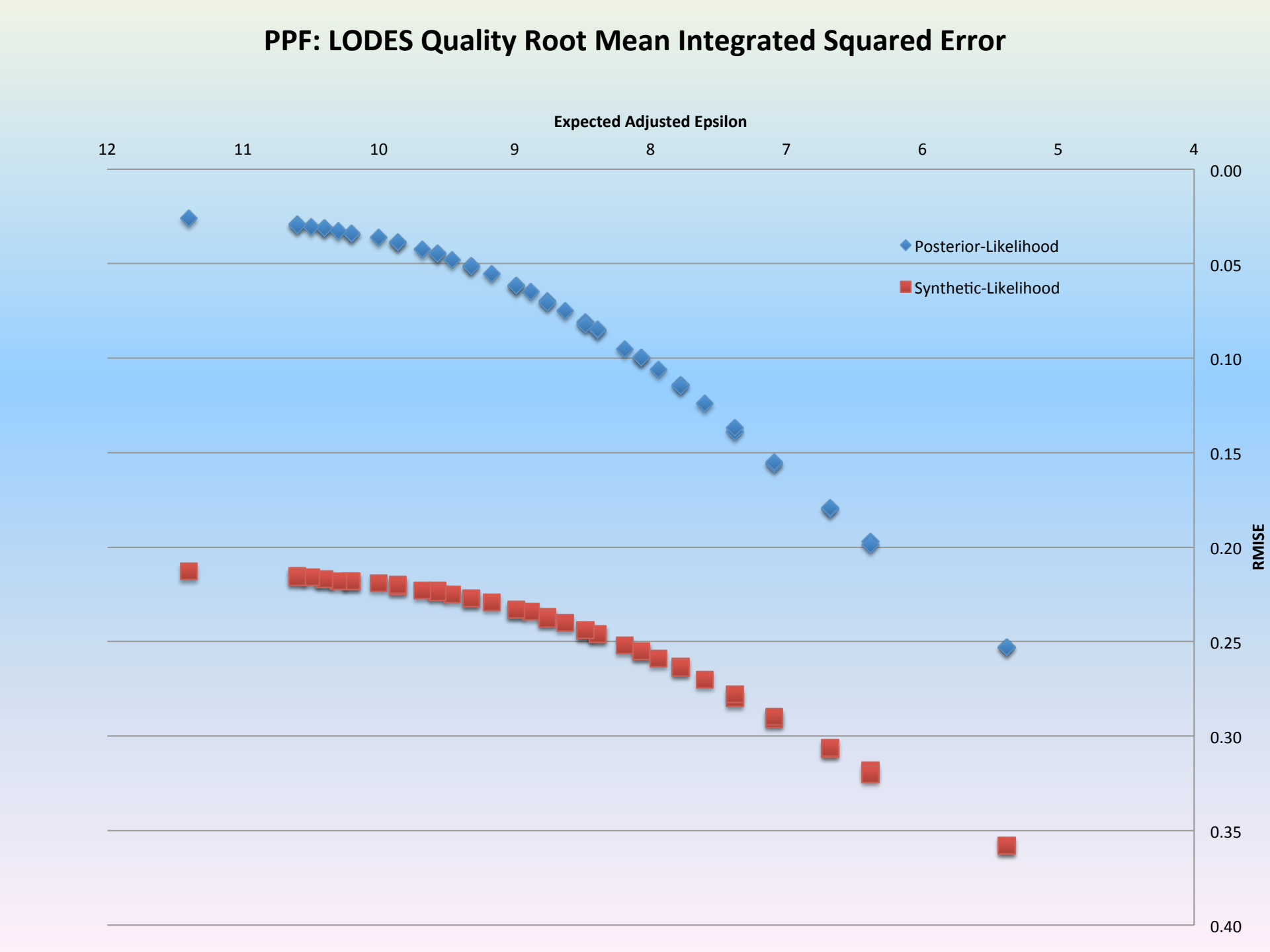

PPF: LODES Quality Measured by Jensen-Shannon Divergence (Zoomed)
Posterior-Likelihood
Synthetic-Likelihood
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1-sqrt(JSD)
12
11
10
9
8
7
6
5
4
3
2
1
0
Expected Adjusted Epsilon

Example 2 PPF

$$I_{RMISE}(\phi) = -\sqrt{\sum_k \left(\hat{\pi}_k(\phi) - \pi_k\right)^2}$$

- Based on the root mean integrated squared error from the same census of N individuals published with privacy ϕ

PPF: LODES Quality Root Mean Integrated Squared Error



Some Implications for Optimal Data Publication/Privacy

- The OnTheMap application at the U.S. Census Bureau published with ϕ = 6.0 attaining data quality of *I* = 0.7
- Using the separable quadratic utility model (specification II) above, this implies that the Bureau considered the ratio of preference covariances to be 0.002, which means that it assumed preferences for information were much more correlated with income than were preferences for privacy.

Alternative Specification

$$SWF: \min_i v_i\left(y_i, I, \phi, p\right)$$

$$PPF: f\left(I, \phi, b\right) = 0$$

- Known as the Rawlsian social welfare function
- Conjecture: differential privacy with $\varepsilon = -\phi$ chosen for the correct marginal individual (the one whose utility is the minimum at the optimum) is the global optimum privacy

Wrapping Up

- I have tried to pose an old problem (public good provision) in a manner that might incite mathematicians to consider models of optimal data production and protection
- This work would build on the existing CS and SDL protection methods by explicitly examining how the protection technology interacts with the data quality measure (PPF), and how preferences interact with the publication choices (SWF)